Introducing Textual Measures of Central Bank Policy-Linkages Using ChatGPT

Lauren Leek[1 2], Simeon Bischl[2], and Maximilian Freier[2]

[1]Department of Political and Social Sciences, European University Institute
[2]DG Economics, European Central Bank

February 19, 2024

Abstract

While institutionally independent, monetary policy-makers do not operate in a vacuum. The policy choices of a central bank are intricately linked to government policies and financial markets. We present novel indices of monetary, fiscal and financial policy-linkages based on central bank communication, namely, speeches by 118 central banks worldwide from 1997 to mid-2023. Our indices measure not only instances of monetary, fiscal or financial dominance but, importantly, also identify communication that aims to coordinate monetary policy with the government and financial markets. To create our indices, we use a Large Language Model (ChatGPT 3.5-0301) and provide transparent prompt-engineering steps, considering both accuracy on the basis of a manually coded dataset as well as efficiency regarding token usage. We also test several model improvements and provide descriptive statistics of the trends of the indices over time and across central banks including correlations with political-economic variables.

Email: laurencaroline.leek@eui.eu

This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB. We thank the Digital Innovation & Infinity team at the ECB for their technical support. We would also like to thank Waltraud Schelkle, Simon Hix and the participants of LSE EU550 for comments on previous versions.

1. Introduction

Monetary policy is intricately connected to fiscal and other government policies, financial market developments and regulation. Since the 1990s, a broad consensus prescribes monetary policy to be delegated to independent central banks to ensure that monetary policy is able to keep inflation under control (Rogoff, 1985). Yet, monetary policy does not operate in a vacuum. First, monetary and fiscal policy are closely linked, given their joint objective of macroeconomic stabilisation (cf. the macroeconomic policy-mix literature, e.g., see Davig and Leeper (2011)). Second, monetary policy is not detached from financial stability considerations. On the one hand, monetary policy is one of the key determinants of financial cycles (Miranda-Agrippino and Rey, 2020). On the other hand, regulation and macro-prudential policy are key determinants for stable financial markets, which are, in turn, a prerequisite for the conduct of a price stability-oriented monetary policy (BIS, 2016).

This paper presents novel indices that identify monetary, fiscal and financial linkages in central bank communication. These policy-linkages can conceptually be differentiated into (i) monetary, fiscal and financial dominance and (ii) monetary-fiscal and monetary-financial coordination. First, monetary dominance is defined as a regime by which government policies accommodate the monetary policy objective of maintaining price stability. Fiscal dominance, however, implies that monetary policy accommodates governments (see Sargent and Wallace (1981)). Accordingly, financial dominance implies that the monetary policy authority subjugates its inflation targets to financial market pressure (see Brunnermeier (2015); Schelkle (2023)). Second, we introduce an additional regime, namely, policy coordination. In this regime, central banks aim to coordinate their monetary policy with the government or financial markets to improve the macro-financial policy mix.

We use a Large Language Model (LLM), specifically, a Generative Pre-Trained Transformer (ChatGPT 3.5-0301), to identify these linkages in speeches of officials from 118 central banks from 1997 to mid-2023. Rather than using a single prompt to classify sentences from central bank speeches, our textual analysis is conducted in coding stages. We optimise this multi-stage classification task using prompt engineering for both accuracy on the basis of a manually coded dataset and efficiency regarding token usage. We provide transparency regarding all the decisions and experiments we conduct for the benefit of future use cases.

Our measure of central bank policy communication brings several improvements vis-a-vis previous studies. First, focusing on speeches helps to overcome a key difficulty in detecting the nature of policy linkages ('dominance' and 'coordination'). More specifically, using text rather than more traditional, macroeconomic data (e.g., Favero and Monacelli (2005); Afonso and Toffano (2013)) or models (Hinterlang and Hollmayr, 2022) allows us to go beyond binary assessment of monetary and fiscal dominance and provides more qualitative information about the interactions. For example, our measure, unlike traditional measures, also captures unconventional monetary policy instruments and their interaction with fiscal or financial policies. Second, our multilevel manually validated and prompt engineered ChatGPT measurement goes beyond word-frequency counts and topic modelling approaches. In contrast to these measures, ChatGPT has a natural understanding of text and can interpret words in their context. This allows us to measure more abstract concepts and relationships previously difficult to detect. Third, the indices also scale beyond human coding of textual data. This is necessary to create indices across time and central banks, especially given the increase of central bank communication in the last two decades (Blinder et al., 2022).

This paper is structured as follows. Section 2 discusses the different forms of interaction between monetary and other government and financial policies – policy coordination and policy dominance – and how these relate to central bank communication. Section 3 discusses the data generating process of our indices, its validation and prompt engineering. Section 4 presents the results of the database describing the trends in the index over time and across central banks. Section 5 concludes.

2. Monetary, fiscal and financial policy-linkages in central bank communication

The central bank independence paradigm posits a strict separation of monetary policy being delegated to a politically independent authority while the government remains in charge of fiscal and other policies (e.g., see Cukierman, Webb, and Neyapti (1992)). Independence is meant to ensure monetary dominance, forcing fiscal or financial authorities to subordinate their policies to the central bank's price stability objective (Sargent and Wallace, 1981). To be more specific, the central bank is expected to control its target variable – typically the short- or medium term HICP – independently, with governments adjusting their policies to the monetary policy objective. Monetary dominance thus implies that monetary policy in principal should not be available for any other policy purpose but to steer the rate of inflation.

Central bank independence is meant to protect the central bank from fiscal dominance. Fiscal dominance describes the situation in which the central bank subordinates its price stability objective to fiscal policy objectives, i.e., the situation in which monetary policy is primarily driven by fiscal considerations rather than maintaining price stability (Sargent and Wallace, 1981). Fiscal dominance may imply outright monetary financing of government debt and intervening directly in the sovereign bond market to stabilise government debt by the central bank. Another example of fiscal dominance is the case that the central bank accommodates the fiscal authority budget deficits. In normal circumstances, the central bank would be required to raise its interest rates to combat these inflationary pressures. However, under a fiscal dominance regime, the central bank does not raise rates in response to these inflation pressures. Low nominal interest rates help the government to sustain its deficit while high inflation erodes the real value of the debt stock (Leeper, 1991).

More recently, scholars have also pointed out that central banks face risks of 'financial dominance' (e.g., Fraga, Goldfajn, and Minella (2003); Brunnermeier (2015); Diessner and Lisi (2020)). Following the logic of fiscal dominance, financial dominance describes a situation in which the central bank is not able or willing to tighten its policy stance as this would threaten the stability of the financial system given the over-leveraged financial system. Brunnermeier (2015) argues that central banks see themselves forced to come to the rescue of financial institutions when facing contagion across banks and a more systemic crisis. This can be driven by concerns about negative feedback effects from the financial markets (such as contagion and doom loops) and central banks may be forced to bail-out or recapitalise banks. In other words, systemic financial risks are pressuring monetary (and fiscal) authorities to ensure the survival of the financial system, subordinating the inflation target to the financial stability objective.

In their seminal piece: "Some Unpleasant Monetarist Arithmetic", Sargent and Wallace (1981) describe monetary and fiscal policy interaction as a competitive game: “Which authority moves first, the monetary authority or the fiscal authority? In other words, who imposes discipline on whom?” (p.7). However, the interaction between the monetary authority and the government does not necessarily resemble Sargent and Wallace’s competitive game. In practice, governments do not continuously and as a matter of principle aim at undermining monetary dominance and exerting fiscal dominance. This allows room for coordination between central banks, governments and financial markets. In this setup, agents benefit from coordinating their activities by making decisions that complement each other and thus maximise the payoffs for all.

We use the term ‘monetary-fiscal coordination’ as referring to the governments and central banks coordinating their policies to offer an optimal or stabilising policy mix and, therefore, accommodate each other in a non-hierarchical manner. For example, in the case of macroeconomic shocks, it may be desirable to coordinate monetary and fiscal policy to ensure optimal output smoothing while keeping inflation at target. This may be especially prevalent when monetary policy approaches the zero lower bound (ZLB) since fiscal policy in that case has to take a larger role in macroeconomic stabilisation. Specifically, to the extent the ZLB reduces the effectiveness of monetary policy relative to fiscal policy in stimulating demand, this may justify governments to raise demand when negative shocks to economic activity occur (Haldane, 2020).

In addition to monetary-fiscal coordination, we introduce the term, ’monetary-financial coordination’ to refer to the situation in which there is coordination of monetary and financial policy and regulation. Central banks are well placed to internalise these complex interactions between monetary and macro-prudential policies since monetary policy affects credit growth which has implications for the health of the financial system (Maddaloni, Mendicino, and Laeven, 2022). A pure monetary dominance regime would require the central bank to obstinately pursue its price stability objective with no regard for financial stability considerations and, in case the financial market bubbles burst, ‘mop up’ afterwards (ECB, 2020). Alternatively, it can coordinate its policies with financial regulators by pushing for adequate capital and liquidity buffers. For example, it can push for stability-oriented financial market regulation, which, in turn, may help prevent financial exuberance and require less forceful and economically potentially less damaging monetary policy interventions.

We suggest to measure these policy-linkages using central bank communication. There has recently been a rise in the use of central bank communications in research, in particular speeches, to determine stances of central banks under high secrecy and limited public information (e.g., Baerg and Lowe (2020); Schonhardt-Bailey (2013); Bennani and Neuenkirch (2017); Moschella and Diodati (2020); Ferrara (2020)). While the literature on monetary policy communication and its predictive power is abundant together with the literature on the (macro)economic effects of central bank communication (e.g., Hansen and McMahon (2016); Swanson (2021)), the literature on central bank communication on fiscal and financial policy is more scarce. The focus in this literature is primarily on measuring the intensity of central banks’ fiscal communication, that is, how much central banks talk about fiscal policy and to quantify the direction (e.g.

“hawkish” vs “dovish”) of that communication (Marozzi, 2021). Heinemann and Kemper (2021), for instance, looks (manually) at whether governors positions and whether board members take hawkish, neutral or dovish positions on fiscal policy. Moreover, Aruoba and Drechsel (2022), building on foundational work of Romer and Romer (2004) show the usefulness of natural language processing by using the documents of the Fed’s Federal Open Market Committee meetings to predict monetary policy shocks. These studies conclude central bank ’talk’ is not necessarily cheap but sets policy directions and can lead to actual responses.

The existing scholarship on monetary policy-interactions in the economic literature mainly focuses on monetary versus fiscal dominance. Most of these studies, rely on economic models with New-Keynesian elements, most often Dynamic Stochastic General Equilibrium (DSGE) models (e.g., Davig and Leeper (2009); Bianchi and Ilut (2017)). These models capture the market interactions of households, firms, the government and the central bank and can be used to conduct counterfactual policy experiments to determine the effect of policy rules on outcomes like inflation. This literature often uses the terminology of Leeper (1991) describing historical periods with either "active" or "passive" fiscal and monetary policy. An active monetary policy regime is as a scenario where the central bank prioritizes the control of inflation through adjustments in nominal interest rates without accommodating fiscal deficits, thus signalling monetary dominance, whereas active fiscal policy is indicative of fiscal dominance. Markov-switching regressions (Favero and Monacelli, 2005) are commonly used to used to estimate regime changes endogenously. These models introduce the ability to capture structural changes in the behavior of economic agents to represent periods of fiscal and monetary dominance by, for example, altering the central bank’s reaction to a deviation from its inflation target. Closely related to this are a number of papers that model dominance by integrating Markov-switching processes with DSGE models (e.g., see Bianchi and Ilut (2017); Davig and Leeper (2011); Chen, Leeper, and Leith (2022)). Other approaches involve testing for Granger causality between fiscal and monetary variables in a vector autoregression setup (VAR) (Sabaté, Escario, and Gadea, 2015), as well as testing for the significance of fiscal variables directly inside policy rules Ahmed, Aizenman, and Jinjarak (2021), such as the Taylor rule. Even more recent approaches use machine learning techniques to classify an unobserved economic state using simulated DSGE data to detect periods of fiscal and monetary dominance (Hinterlang and Hollmayr, 2022).

We argue that using a central bank communication approach can complement and even improve on the existing measurement approaches. First, existing approaches typically focus on fiscal deficits and the policy rate of the central bank to identify monetary or fiscal policy regimes. However, these indicators can be influenced by many other variables and do not take account of the rapid expansion of non-standard monetary policy tools such as quantitative easing and monetary policy transmission safeguards. In this context, central banks also resort to communication as a policy-making tool in itself, namely, through making statements about the likely future path of interest rates (forward guidance).[1] Second, these models rely heavily on simplifying assumptions about the economy, such as a limited number of distinct actors and perfect foresight (e.g., see Stiglitz (2017) for more information). Traditional DSGE models often do not consider the financial sector at all or have a simplified representation of the financial sector which does not capture the impact of the pressure from financial markets. Third and most importantly, in the existing literature on monetary policy interactions, coordination as a form of policy-linkage is overlooked. Coordination is impossible to detect using DSGE or related models. However, policy-linkages between independent central banks and other policy agents are not only described my monetary, fiscal or financial dominance, but also by coordination between these three policies.

3. Constructing and validating our indices

In this section, we will first describe our rationale for our method of constructing our indices using a Large Language Model (LLM), in our case ChatGPT 3.5. Second, we will describe our textual corpus and pre-processing steps. Third, we will describe our multi-level coding scheme and provide examples. Fourth, we present the validation of our indices including our usage of prompt engineering and decisions made regarding token efficiency. We also test a range of other models. Fifth and last, we will discuss how we construct our policy-linkage indices from our textual corpus.

3.1. Choice of ChatGPT

The purpose of this paper is to develop indices which identify policy linkages in central bank communication. These indices not only identify policy linkages but also determine their nature (policy coordination or dominance) over time and across central

[1] Forward guidance is in particular used if the central bank can no longer cut policy rates because they already have reached their zero lower bound. Via forward guidance, central banks provide additional information regarding their likely response to economic developments, which can anchor expectations about future policy rates and reduce uncertainty (Strasser et al., 2019).

banks. Identifying these patterns from text is a complex reasoning task that requires substantial domain knowledge. It is typically done through manual coding. However, given the scale of this project – more than 18000 speeches in which there are more than 2 million sentences which need to be coded – this is an unrealistic avenue.[2] An automated procedure is therefore necessary.

Traditional text-as-data methods – for example, counting the occurrence of certain words or determining the topic of a speech using topic modelling (e.g., using bag of word or word embedding methods) are not suited to determine the often subtle portrayals of policy dominance and coordination. This can be illustrated by the following excerpt: "Accordingly, and as recently confirmed by the ECOFIN Council and the European Council, the fiscal policy provisions of the Maastricht Treaty and the Stability and Growth Pact should continue to be applied fully. The fiscal rules are one of the indispensable pillars of EMU and the single currency, which must remain firmly in place so as not to undermine the confidence in fiscal sustainability. Finally, the current situation calls for ensuring the high quality and timeliness of statistical information on government interventions to ensure the transparent and accountable use of public funds." from an ECB Press Conference by Jean-Claude Trichet on November 6 2008, which discusses fiscal rules that protect the central bank from fiscal dominance and are thus imply monetary dominance. ChatGPT classifies this excerpt correctly, while an algorithm without contextual understanding would not be able to identify the implication for monetary policy contained in this statement. We overcome these two empirical challenges using a (close) to state of the art large language model, namely, ChatGPT 3.5.[3]

ChatGPT functions as a "zero-shot learner". This means it can perform classification tasks without requiring additional training or fine-tuning. It produces its answers based on the instructions and input text contained in the prompt and solely relies on the knowledge and instructions following capabilities built into the model. The crucial advantage of this is that we can conduct our analysis without requiring a substantial manually labelled dataset as training sample. Moreover, ChatGPT has demonstrated strong performance across a number of natural language processing tasks in zero-shot learning setups (Laskar et al., 2023). ChatGPT's capabilities are based on a Generative

[2] During our manual classification of our validation set, it took us around one minute per sentence. Thus, given the sample of two million sentences, it would take around 33,000 hours or almost 4 years cumulative to classify the entire dataset manually.

[3] We use version gpt-3.5-turbo-0301. At the time of writing this paper ChatGPT4 was not yet available to us, substantially more costly and subject to restrictive rate limits. Similarly, Google's Gemini was not yet released when the main analysis was run. We do run our validation set with ChatGPT4 and Gemini Pro.

Pre-trained Transformer (GPT) model.[4] ChatGPT is fine-tuned to closely follow instructions using a process called reinforcement learning from human feedback (RLHF). This method enhances the GPT model by having humans review its responses. Based on their evaluations, the model's parameters are adjusted to generate more desirable responses (Ouyang et al., 2022). The prompt-following ability makes ChatGPT easier to use than next word prediction models and readily applicable for classification tasks (e.g., see Hansen and Kazinnik (2023) which is the latest and so far to our knowledge only paper applied to central bank communication).

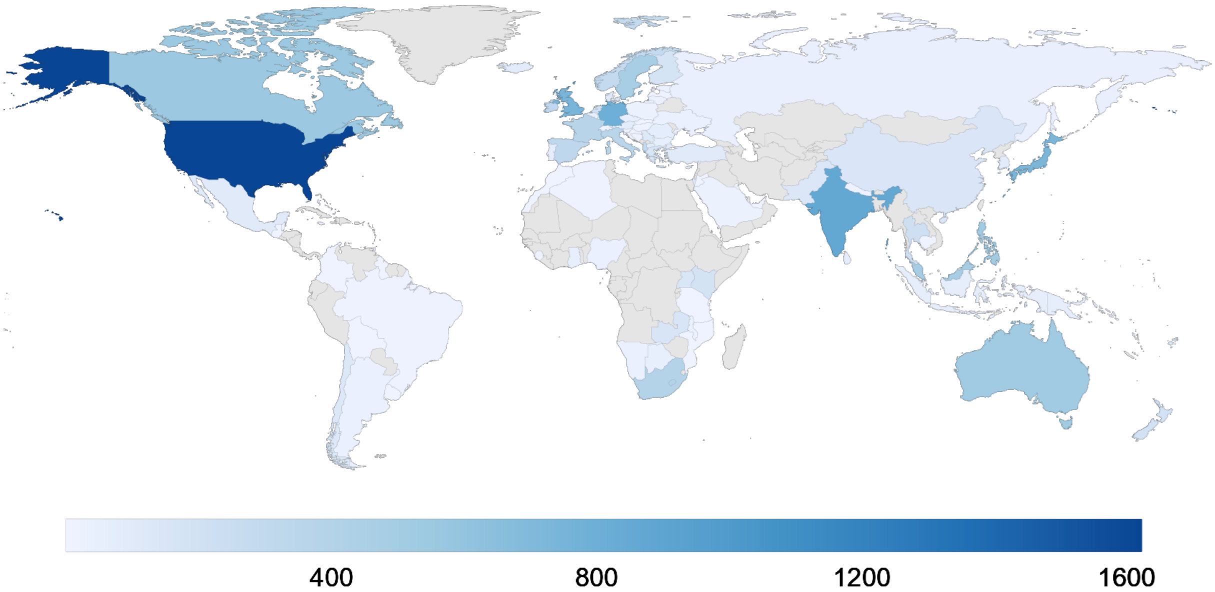
3.2. Corpus and pre-processing

Our corpus of central bank communication is scraped from the Bank of International Settlement website. It covers mostly speeches, but also press conferences and interviews of central bank officials from from 118 central banks and monetary institutions over the period from January 1997 to July 2023. This dataset totals around 18,000 documents (see Figure 1 for the geographic distribution of the frequency of speeches and see the Appendix Table A1 for more details on the speeches). The speeches are downloaded in PDF format from the BIS website, converted to text and cleaned using various pre-processing steps following standard methodology (e.g., removing page numbers, footnotes headers, chart titles, new page characters, URLs, headers and line breaks). Subsequently, we break up the speeches into single sentences[5] We also remove anything less than 6 tokens, anything more than 200 tokens, less than 20 characters, sequences of whitespace characters and all sentences consisting of less than two thirds ASCII characters. These are conservative heuristics to remove text that is very unlikely to contain relevant information. We retain information on the ordering which will be used in further steps.

[4] GPT belongs to a family of transformer models which use the self-attention mechanism developed by Vaswani et al. (2017). The self-attention mechanism is well parallelisable in training and is therefore more scalable than the previously used recurrent neural networks, allowing for larger neural networks with more parameters (Wolf et al., 2020).

[5] Since the source PDFs do not follow an entirely standardized layout, information about paragraphs cannot be retained. Also owing to the PDF format, we cannot fully clean the content of the speeches; thus, infrequent footnotes, chart annotations and citations remain. We rely on ChatGPT to classify these as “other” and “none”, which are not considered for our final indicators.

FIGURE 1. Geographical spread of number of speeches available per central bank



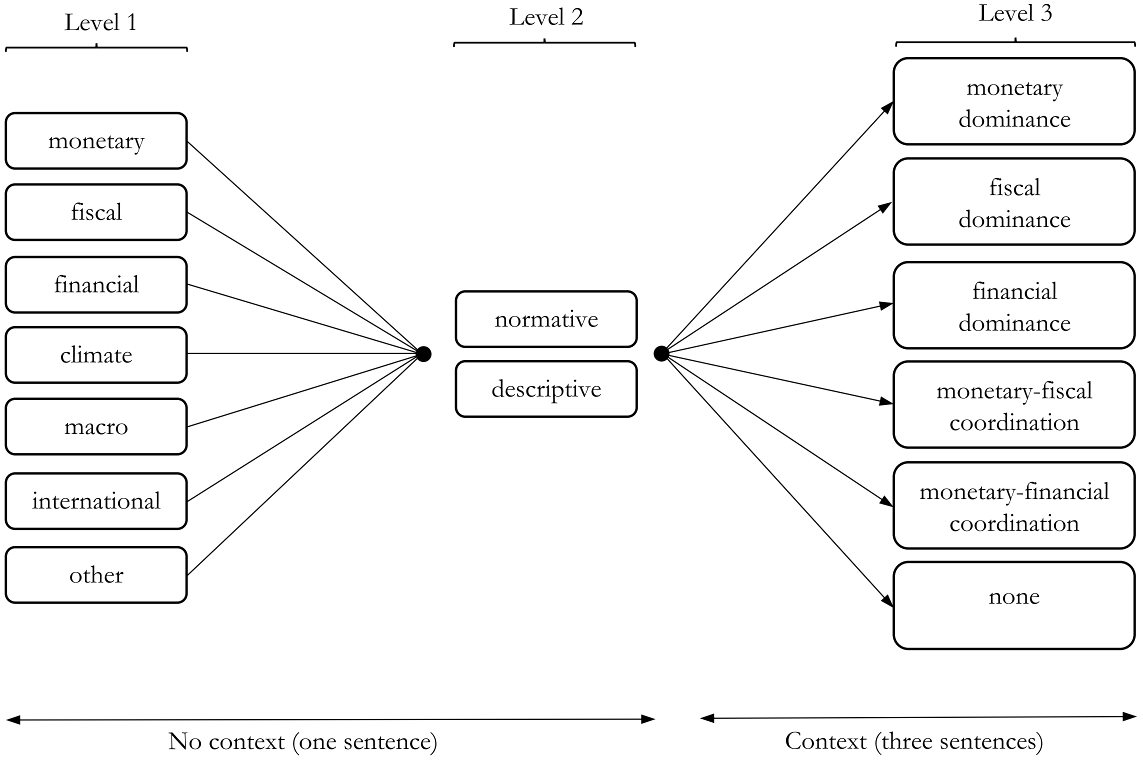
Note: Color scale indicates the number of speeches per country contained in our dataset. See our Appendix Table A1 for the full list of central banks and the number of speeches per central bank. The central bank with the most speeches is the ECB with 2377 speeches (not shown on the map in order to be able to see which National Central Banks within the Eurosystem are included).

3.3. Identifying policy linkages

Our classification scheme consists of three steps. In the first stage (level 1), we determine the topic of the single sentence: “monetary”, “fiscal”, “financial”, “climate”, “macro”, “international” or “other”. We determined these topics by inductively examining 100 randomly drawn sentences from the entire corpus. Although there are many more topics (e.g., see Hansson (2021)), the first level mainly aims at examining the frequency of topics addressed by central banks. Including also other policy topics creates a database for further analyses (e.g., to examine whether mentioning certain topics is highly correlated with dominance and whether certain topics more often result in dominance or coordination). In the second step (level 2), we determine whether the sentence is descriptive or normative. That is, we prompt ChatGPT to determine whether the sentence simply describes monetary, fiscal or financial developments or policies or offers a value judgement. Very often central bankers use descriptive sentences to portray statistical information such as inflation or growth rates. Alternatively, the sentence can be classified as normative if it prescribes a certain policy action, be it to the central bank itself or to other institutions. Again, the data from this stage can be used for further analysis, namely, are certain topics or more normative or are there certain central banks that talk more normative while other central banks talk more factual? In the third step (level 3), we classify the nature of the policy-linkages. To operationalise the definitions of monetary, financial and fiscal dominance and coordination, we rely on an actor-centred approach. Dominance places one actor hierarchically above another actor (e.g., in the case of fiscal dominance, the monetary authorities accommodate the fiscal authorities). In the case of coordination the actors are placed on an equal footing.

In contrast to levels 1 and 2, level 3 categories are less clear cut, generally require more complex reasoning steps and often depend on contextual information. For instance, accurately interpreting the intent behind a speech often requires considering both a factual statement and a subsequent value judgement together, as only their combination provides a clear understanding of the speakers intentions. Therefore, we add the sentence before and after to classify 3-sentence excerpts in level 3 only. We do this for level 3 only since it is computationally intensive as it effectively triples the text that needs to be classified. Figure 2 below summarises the three level coding scheme.

FIGURE 2. Three stage coding scheme GPT to identify dominance and coordination.



To illustrate the various forms of policy linkages we obtain with the level 3 classification, Table 1 provides a number of examples and brief explanations. These sentences are taken from a 1000 sentence random sample created to validate our ChatGPT results. See the Appendix sections C.2 and C.3 for guidelines to classify ambiguous sentences and examples of all categories of level 1 and 2.

TABLE 1. Classification examples

Classification	Example
Monetary Dominance	"Furthermore, monetary policy implementation in line with the market efficiency principle would need to remain without prejudice to our primary mandate of safeguarding price stability." (Retrieved from: The European Central Bank, 14-06-2021). *Explanation:* the topic concerns a monetary topic and they emphasize their primary mandate of price stability being above other priorities. Therefore, this sentence can be classified as monetary dominance.
Fiscal dominance	"Moreover, although most of the resources administered by the BIS are invested in financial assets of top quality at international level and their exposure to the various risks are managed conservatively, a greater portion of such funds could be spend toward the direct purchase of debt denominated in local currencies of emerging countries or to the use of them as collateral of certain bond issuance of countries with limited depth of their financing markets in local currency." (Retrieved from the Central Bank of Argentina, 09-07-2008.) *Explanation:* This sentence refers to funds being spend towards the direct purchase of debt (=monetary financing) instead of considering pure price stability considerations, thus we consider this sentence to be fiscal dominance.
Financial dominance	"It is thus significant that our flexible and abundant provision of liquidity contained market participants' concerns over liquidity financing." (Retrieved from the Bank of Japan, 04-07-2002) *Explanation:* This sentence states that monetary policy is accommodating financial markets by providing liquidity, thus showing that financial markets are a consideration for the bank in conducting their monetary policy.
Monetary-fiscal coordination	"Since restarting our strategy review, we have introduced a new work stream on monetary-fiscal interactions precisely to address such questions." (Retrieved from the European Central Bank, 30-09-2020). *Explanation:* This sentence refers to the monetary-fiscal interactions which is a key policy in the monetary-fiscal coordination.
Monetary-financial coordination	"If market participants are willing to continue to work together, then we can safely achieve the transitions needed to create a better and more robust system that will help to ensure our ongoing financial stability." (Retrieved from the Board of Governors of the Federal Reserve System, 07-11-2017). *Explanation:* this sentence shows that the bank wants coordinate with market participants to ensure financial stability.

3.4. Validation and prompt engineering

Employing ChatGPT as our classifier requires some efforts in prompt engineering. In contrast to supervised machine learning algorithms, there is no training dataset and hardly any tuneable parameters to optimise. The behaviour of the model and thereby also the accuracy with which it can classify sentences is entirely determined by how it is instructed, i.e. the prompt that is given to the model. At the time of writing, there are no established best practices yet on how to write an optimal prompt. Academic contributions that analyse the trade-offs in designing prompts do not exist to the best of our knowledge yet.

The aim of this section is to describe and offer guidelines on the steps and experiments in determining the optimal prompt for our use case. This may also serve as a reference for future research. Since there are an almost infinite number of prompt variations, it is not feasible to test every possible modification to the prompt. Systematically testing prompts is further complicated by the fact that ChatGPT is not entirely deterministic. ChatGPT occasionally hallucinates output categories or does not follow the prescribed output format, which requires manual intervention to calculate accuracy metrics. Finding our final prompt was thus an iterative process of making incremental changes to the prompt until further modification no longer resulted in substantial improvements in terms of output stability; that is the models adherence to instructions, and accuracy vis-a-vis a manually validated dataset which we describe below.

3.4.1. Manual validation

Our first step to evaluate different prompts is to manually classify a validation sample, that we treat as ground truth and compare ChatGPT's results against. We determined the quality of a prompt based on a number of validation metrics. Our manual validation sample consists of 1000 sentences that were randomly drawn from the entire sample of two million sentences, making the validation sample representative of the whole dataset with regard to the frequency of each category. All three levels were coded by the three authors independently, using the definitions provided in the prompts and the ambiguous sentence guidelines presented in the Appendix. All three coders coded the same first 400 sentences for two reasons. First, in this way we can calculate coder reliability scores to determine how consistent the assigned labels are across coders. This is especially important given the high level of judgement and abstraction needed to classify some of the sentences. Second, we can use this information to examine how the accuracy and uncertainty in ChatGPT's classification correlates with agreement among human coders. All three coders subsequently coded 200 extra sentences each, expanding our validation set to 1000 sentences. We validate our model using both the full sample and a agreed sample, where all human coders made the same assessment.

Agreement among the independent coders is high. The inter-coder reliability score measures the consistency of coding results (see O'Connor and Joffe (2020) for more information). The average correspondence in coding for level 1 is 84%, 81% for level 2 and 76% for level 3 (see the Appendix section C.4 for additional coder reliability metrics). The slightly lower (but still considerably high) score for level 3 indicates that the classification task is difficult even for human coders trained as (political) economists. Most disagreement in the manual classification comes from the distinction between the coordination and dominance categories, i.e., whether the relationship between the actors is hierarchical or not. In case of disagreement between coders, a majority rule is used to determine the "correct" label.[6]

We used a total of 5 validation metrics to determine the best prompts to run on the entire corpus. These are standard practice and calculated using the scikit-learn package in Python (see Pedregosa et al. (2011) and the Appendix section C.5 for a description of these measures). When determining the best prompt, we pay particular attention to the F1 macro score. This score is sensitive to changes in prediction accuracy in the less frequent categories. This is important given that we are especially interested in being able to identify specific kinds of dominance, particularly fiscal and financial dominance which are relatively infrequent.

3.4.2. Prompt engineering steps

In a second step, we designed and optimised the prompt, that is, the instructions given to ChatGPT. This prompt engineering exercise had three dimensions, namely (i) which and how much information the prompt should optimally contain, (ii) the optimal temperature settings and (iii) the optimal sentence count within a prompt (also considering efficiency). Subsequently, we tried several of the techniques outlined in OpenAI's prompt engineering guidelines.[7]

The first and seemingly most critical choice is what information to provide ChatGPT

[6] In case there is no majority, i.e., all three coders classified the sentence differently, the assigned label is randomly chosen. This affects 5 sentences on level 1 and 7 sentences on level 3. The binary classification of level 2 ensures that there is always a majority label.

[7] See https://platform.openai.com/docs/guides/prompt-engineering for the guidelines.

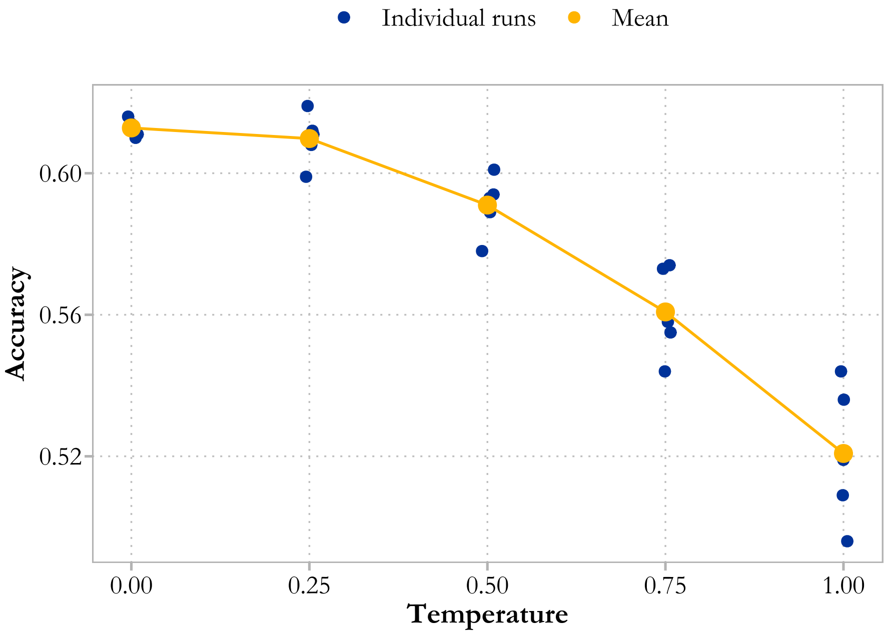
regarding the classification task it should conduct. Do we need to give exact definitions and examples of, for instance, fiscal dominance? Or can we rely on ChatGPT's understanding of what fiscal dominance is, encoded by its billions of parameters?[8] However, correctly defining concepts does not imply that ChatGPT can operationalise these to classify sentences. Our experiments suggest a middle ground – between providing no extra information and very detailed instructions and examples – works best for our classification of level 3. Both a prompt with minimal instructions as well as a more elaborate prompt perform worse than our optimal prompt (see Table A9 and section F.2 in the Appendix for validation metrics and the prompt variations). This indicates that initially providing more information can aid the model, but an excess of details might hinder its ability to differentiate between the main instructions and additional information. Performance does not increase through including overly specific instructions to correct infrequent mistakes. Providing the model with example sentences does improve performance, but as evidenced by the lower accuracy of our more elaborate prompt, the information that is presented to the model should be chosen carefully. Overall, prompts should be concise, but elaborate enough to provide essential instructions.

Second, temperature settings need to be chosen for the prompt. The responses from ChatGPT are never fully deterministic. The degree of variation in the responses can, to a certain extent, be controlled with the temperature settings. A higher temperature leads to a less stable, but more diverse and creative responses from ChatGPT. To validate the temperature setting, we ran temperature levels between 0 and 1 with 0.25 increments in the temperature five times on the validation set. As expected, the accuracy, measured as overlap between the human and ChatGPT classification, varies more across runs as temperature increases (see Figure 3). Average accuracy is very similar for temperature settings smaller or equal to 0.5. Since choosing the lowest temperature settings does not seem to incur a cost in terms of average accuracy, we select a temperature of zero. Higher stability of results across runs facilitates the prompt engineering process as less runs are required to determine the effect of changes to the prompt. Also, lower temperature is helpful for replicability, although perfect replicability is currently not possible with ChatGPT. Typically, more than 97-99% of sentences are classified identically between

[8] When asked for a definition of "Fiscal Dominance" ChatGPT replies accurately. More specifically, ChatGPT4 replied to "Can you define the concept of fiscal dominance in three sentences?" with "Fiscal dominance occurs when a government's fiscal policy—particularly its borrowing and spending habits—dictates monetary policy decisions. In such a scenario, the central bank may be pressured to finance government deficits, often by purchasing government securities or through other means that can lead to inflation. This situation typically arises when a government's debt levels are so high that normal market mechanisms for debt financing are no longer effective, forcing the central bank to step in."

two runs with temperature 0 directly after each other.[9]. In addition, server-side changes to the model configuration are outside of our control and can potentially affect the classification.[10]

FIGURE 3. Variation in accuracy with different temperature settings on level 3 classification.



Note: Every temperature setting was run five times. Blue dots show the accuracy of each run. Yellow dots indicate the average of each temperature setting.

A third consideration is how many sentences to add within a single prompt. We let this choice depend on two criteria: accuracy and efficiency. The effective use of tokens is of importance since the utilization of commercially operated LLMs is billed by token usage and can quickly become cost-intensive, especially given the size of our dataset.[11] LLMs like ChatGPT operate on tokens instead of words whereby on average a words consists of around 1.3-1.5 tokens. Increasing the number of sentences per prompt reduces the number of tokens that are required to classify the entire sample as the

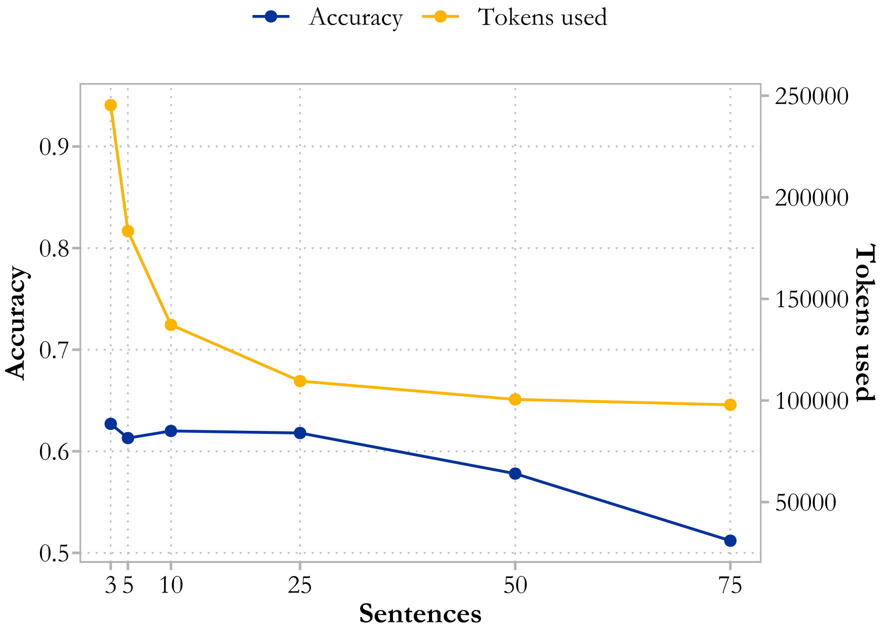
[9] OpenAI introduced a option to set a seed as a beta feature in November 2023. However, passing the seed to the API does not seem to have an effect with our version of the model.

[10] Since November 2023, the API returns a system fingerprint that can be used to trace when changes occur. However, it is still unclear what was changed and whether this affects model behavior. Complete replication of results is only possible when there is unrestricted access to the model, similar to what is available with certain open-source large language models.

[11] Our final level 3 classification used 2,791,384,568 tokens. Increasing this number by even small factor would lead to a substantial rise in expenses.

instructions need to be restated fewer times. The theoretical maximum is given by the maximum context length, that is the number of input tokens a LLM can process. In the case of ChatGPT 3.5-0301, the context length is 4096 tokens, which is equivalent to roughly 100 sentences plus instructions. However, as the sentence count increases, token savings decline quickly since the instructions only make up a small share of the tokens compared to the sentences to classify. Accuracy is highest when classifying 3 sentences inside a single prompt (see Figure 4 for level 3). Including 5 or 10 sentences reduces accuracy slightly, while significantly lowering the token count. Including 10 or 25 sentences inside a prompt yields more accurate results than 5 sentences. A possible explanation is that showing multiple sentences at the same time, helps ChatGPT to distinguish the categories. On the other hand, if the number of sentences gets too large, its ability to correctly classify sentences decreases. We find a similar bell-shaped relation for level 1 and level 2 (see Appendix section D). Taken these points into consideration, we opt for 10 sentences per prompt on all three levels.

FIGURE 4. Relationship between sentence count, accuracy and token usage



Note: The yellow (blue) line illustrates how token usage (accuracy) varies with the number of sentences that are included in a single prompt.

Lastly, we test the OpenAI prompt engineering guide. The main takeaway from OpenAI's prompt engineering guide is to conduct systematic testing to the degree the unstable output of ChatGPTs allows for it. Our final prompt went through many iterations testing numerous of the suggestions from OpenAI's prompt engineering guide. We find that changing the system message – which should govern the overall behaviour of the model – has little impact (see the Appendix Table A8 for the results of runs with different system prompts). Moreover, clearly structuring the prompt with delimiters helps to stabilise the model output and instructions regarding the output format are best placed at the end of the prompt.[12] The final prompts for levels 1 to 3 are given in the below.

[12] We did not experiment with techniques that elicit a reasoning process in the model output like chain-of-thought prompting (see, Wei et al. (2023)) as this would increase the output tokens by at least 50 times compared to directly prompting for the label.

TABLE 2. Final Prompts

Level 1

Classify sentences from a central bank speech as one of the following categories

Monetary if:
•The sentence addresses monetary issues (e.g., inflation, price stability, primary mandate, interest rate)

Fiscal if:
•The sentence addresses fiscal issues (e.g., sovereign debt, budget balance, fiscal governance, taxes, pensions)

Financial if:
•The sentence addresses financial issues (e.g., banking supervision, financial instability, credit risks)

Climate if:
•The sentence addresses climate issues (e.g., environmental issues, CO2, climate change, sustainable development goals)

Macro if:
•The sentence addresses macroeconomic issues (e.g., GDP, economic growth, unemployment, productivity, economic outlook)

International if:
•The sentence addresses international economics issues (e.g., trade, exchange rates, capital mobility, tariffs)

Other if:
•The sentence does not relate to the topics of monetary, fiscal, financial, climate, macro or international economics.
•This category should be used as the default category
Reply only with the number of the sentence and the assigned label. These are the sentences: . . .

Level 2

Classify sentences from a central bank speech as either normative (value judgement) or descriptive. These are the sentences: . . .

TABLE 2. Final Prompts *(continued)*

Level 3

Please classify excerpts from central banker speeches in one of the following categories:

•“none” if there is no reference to monetary, financial or fiscal developments

•“none” if the excerpt describes monetary, financial or fiscal developments, that is, if the speaker does not make any normative reference to monetary, financial or fiscal policy (example sentence: “the deficit is expected to reach 2.5% of GDP in 2020”)

•“monetary dominance” if the excerpt clearly and explicitly says that the central bank subordinates fiscal or financial policies to the central bank’s monetary policy objective of price stability (example sentence: “the role of the central bank is not to ensure financial stability or fiscal sustainability but to maintain price stability”)

•“monetary-financial coordination” if the excerpt suggests that the central bank and financial regulators should cooperate, this is, where the speaker says that monetary policy and financial regulation are best coordinated to achieve the right policy mix (example sentence: “higher capital requirements will increase the resilience of the banking system and support the transmission of monetary policy”)

•“monetary-fiscal coordination” if the excerpt suggests that fiscal authorities and the central bank should cooperate, this is, where the speaker says that monetary and fiscal policy are best coordinated to achieve the right policy mix (example sentence: “the deficit should remain below 3% of GDP not to further increase inflationary pressures”)

•“financial dominance” if the excerpt clearly and explicitly says that the central bank subordinates to financial markets or the financial regulation authorities, that is, where the speaker says that monetary policy is primarily driven by financial stability considerations rather than maintaining price stability (example sentence: “lower interest rates will ensure the stability of the banking system”)

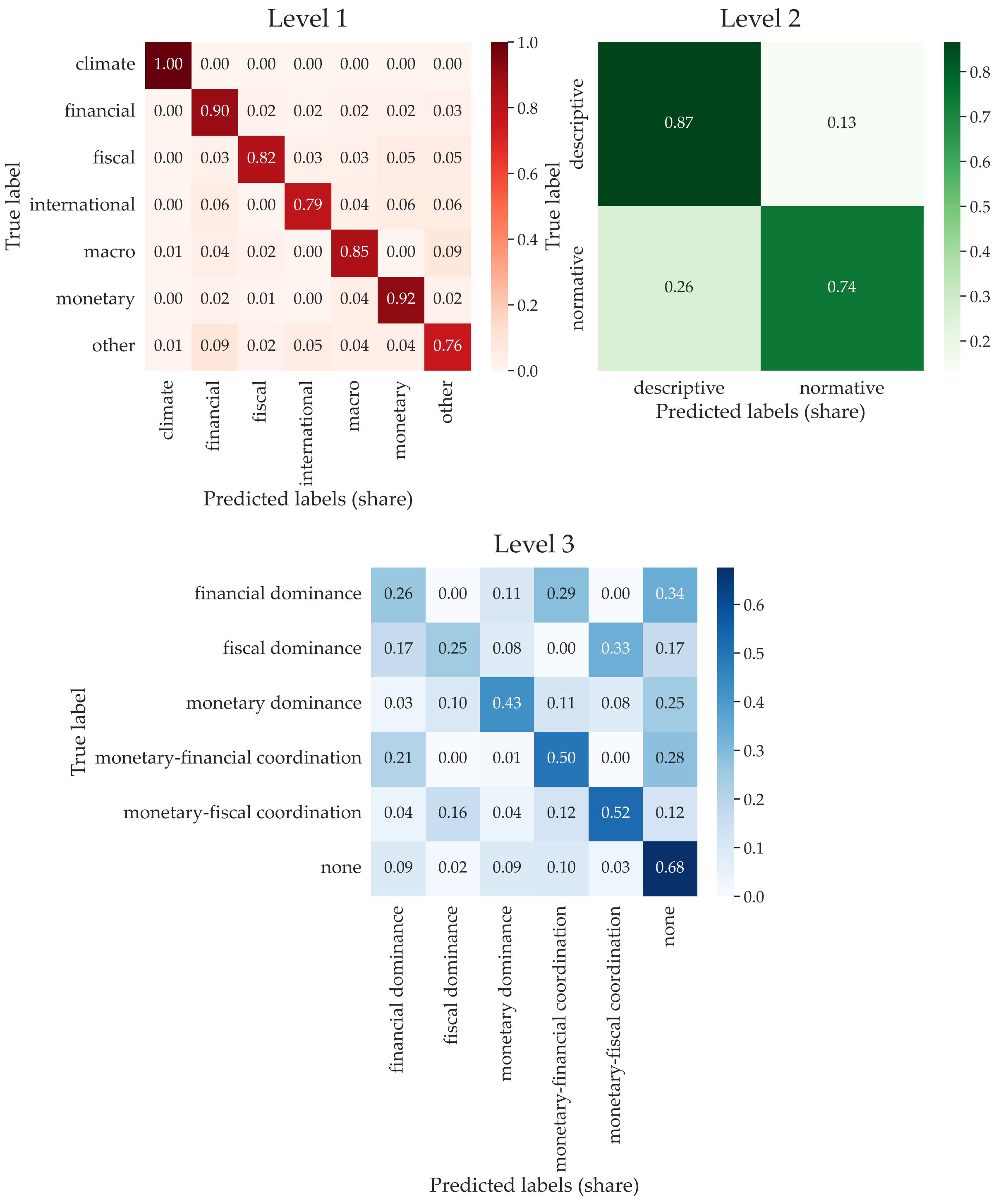
•“fiscal dominance” if the excerpt clearly and explicitly says that the central bank subordinates itself to fiscal authorities, that is, where the speaker says that monetary policy is primarily driven by fiscal considerations rather than maintaining price stability (example sentence: “lower interest rates will ensure that public finances remain sustainable”)

Classify each of the excerpts individually. Reply only with the number of the excerpts and the assigned label. These are the excerpts: . . .

3.4.3. Validation scores

Figure 5 below, shows the confusion matrices, i.e., a matrix visualisation of of the distribution of predicted labels per category. The confusion matrices can be used to determine which classes are most frequently confused with each other, which classes the model predicts most accurately, and where the model's weaknesses lie in distinguishing between classes. Taking the human-coded classification as the ground truth, ChatGPT assigns the correct label with the highest probability across all three levels with the exception of financial and fiscal dominance on level 3. In our level 3 classification, the the most frequent mistake is always the related dominance/coordination category except for financial dominance which is most often confused with "none". (e.g., a sentence that was classified as monetary-fiscal coordination in the validation set has a 52% probability to be correctly classified and a 33% probability to be classified as fiscal dominance). Looking at the excerpts that should not be categorized under dominance or coordination, we find that our ChatGPT classifier has a tendency to over-represent financial dominance and coordination. This could be explained by the fact that financial dominance and coordination are not as thoroughly discussed in existing literature compared to the monetary and fiscal counterparts, which likely means that these concepts are less represented in ChatGPT's training data, reducing its ability to accurately identify dominance and coordination with regard to financial markets. With regards to our 5 validation metrics, level 1 and 2 perform well in all metrics with at least 83% of sentences matching the human classification, and F1 scores of close to or above 80% (see Table 3) in the full sample. ChatGPT's ability to mimic the human coding is lower for level 3 but can still be considered relatively high given the complexity of the task, the room for interpretation, and the degree of disagreement also present among human coders. The accuracy in classification at Level 3 is greater in the sample where human coders agree compared to the entirety of the dataset. For Levels 1 and 2, however, the discrepancies in classification between the agreed sample and the full dataset are small.

FIGURE 5. Confusion Matrix (based on our full validation sample)



Note: The confusion matrices display the correspondence between actual categories and predicted labels. For instance, among the sentences categorized as 'financial' by the human coders in the validation set, 90% were accurately labeled by ChatGPT. The remaining 10% were misclassified, with their distribution being relatively uniform across other categories, with the exception of the climate category.

TABLE 3. Validation metrics of final model on full and agreed sample

	Accuracy	F1 (weighted)	F1 (macro)	Precision (macro)	Recall (macro)
B. Agreement sample					
Level 1	0.83	0.83	0.78	0.75	0.84
Level 2	0.83	0.84	0.75	0.74	0.77
Level 3	0.68	0.74	0.35	0.32	0.46
B. Full sample					
Level 1	0.85	0.85	0.83	0.80	0.86
Level 2	0.83	0.84	0.79	0.78	0.80
Level 3	0.62	0.66	0.36	0.34	0.44

Note: The agreed sample includes 316 sentences for level 1, 295 sentences for level 2, and 307 sentences for level 3. For inclusion in the agreement sample, all coders must have coded identically on the respective level. Precision and recall are macro averages taking the unweighted average of the precision and recall score of all categories.

3.4.4. Improvements and uncertainty measures

In the following, we examine a range of other LLMs, which include more recent versions of ChatGPT, a fine-tuned ChatGPT variant, and Google's Gemini Pro. We begin by comparing our main ChatGPT model (gpt-3.5-turbo-0301) to the most recent iterations of GPT-3.5 (gpt-3.5-turbo-0125) and GPT-4 (gpt-4-turbo-0125). We use the same validation set and the prompt we engineered for our main model. Overall, we find considerable improvements, especially when using GPT4 for level 3. Note that the validation metrics in Table 3 do not always move in the same direction. When given the same prompt, GPT-4 is much less likely to classify in any of the dominance categories, aligning more closely with human classification. As a result, we find much higher precision and a somewhat lower recall for GPT-4. We also noticed that GPT-4 always replied with one of our predefined labels, which makes evaluating prompts much easier. This illustrates that the choice of model plays a more significant role in influencing validation metrics compared to incremental refinements of the prompt. Due to token limits on the runs in the Open-AI platform and due to GPT-4 being at least 10 times the cost per token, we still opt for the older GPT3-based version.

TABLE 4. Validation metrics for ChatGPT 3.5-0301, 3.5-0125, 4-0125

	Level 1			Level 2			Level 3		
	3.5-0301	3.5-0125	4-0125	3.5-0301	3.5-0125	4-0125	3.5-0301	3.5-0125	4-0125
A. Agreement sample									
Accuracy	0.79	0.83	0.79	0.85	0.83	0.84	0.75	0.68	0.86
F1 (weighted)	0.79	0.83	0.79	0.85	0.84	0.85	0.79	0.74	0.85
F1 (macro)	0.77	0.78	0.75	0.76	0.75	0.78	0.40	0.35	0.39
Precision	0.77	0.75	0.73	0.77	0.74	0.76	0.37	0.32	0.40
Recall	0.78	0.84	0.81	0.76	0.77	0.80	0.58	0.46	0.41
B. Full Sample									
Accuracy	0.79	0.85	0.80	0.82	0.84	0.82	0.65	0.62	0.78
F1 (weighted)	0.79	0.85	0.81	0.82	0.84	0.83	0.68	0.66	0.76
F1 (macro)	0.77	0.83	0.77	0.75	0.79	0.78	0.38	0.36	0.43
Precision	0.78	0.80	0.78	0.77	0.78	0.76	0.36	0.34	0.59
Recall	0.76	0.86	0.78	0.74	0.80	0.80	0.46	0.44	0.41

Note: The agreed sample includes 316 sentences for level 1, 295 sentences for level 2, and 307 sentences for level 3. For inclusion in the agreement sample, all coders must have coded identically on the respective level. Precision and recall are macro averages taking the unweighted average of the precision and recall score of all categories

Next, we explore fine-tuning and replacing ChatGPT with a competing model (Gemini developed by Google). For the exercise of fine-tuning, we split our sample of 1000 manually annotated sentences into a training sample of 300 sentences and 700 sentences for evaluation. Fine-tuning involves further training of the model using input prompts and the expected outputs. Utilizing the OpenAI interface with default settings, we fine-tune gpt-3.5-turbo-1106 using 60 samples presenting 5 sentences in each sample with the same prompt as before. The resulting model improves significantly upon the vanilla (i.e., not finetuned) gpt-3.5-turbo model and exceeds GPT-4 on the F1-macro score. After fine-tuning, gpt-3.5-turbo becomes much less likely to classify the infrequent dominance categories, while retaining relatively high recall on dominance and coordination, making it the overall best performing model. While fine-tuning yields significant enhancements, we did not opt for it since it also markedly increases the cost of running inference. Running prompts on a fine-tuned chat-gpt-3.5-turbo incurs 6 times the cost per input token compared to the standard gpt-3.5-turbo as of February 2024.

TABLE 5. Fine-tuning, Few Shot learning and Gemini

	ChatGPT			Gemini	
	gpt-3.5-0301	gpt-3.5-fine-tune	gpt-4-1106	Gemini Pro	Gemini Pro Few Shot
Accuracy	0.64	0.77	0.79	0.78	0.79
F1 (weighted)	0.69	0.78	0.78	0.73	0.75
F1 (macro)	0.35	0.43	0.40	0.36	0.40
Precision (macro)	0.33	0.40	0.48	0.44	0.50
Recall (macro)	0.43	0.49	0.40	0.34	0.36

Note: gpt-3.5-fine-tune is a fine-tune on top of gpt-3.5-turbo-1106 using 300 sentences for fine-tuning. The remaining 700 sentences of the validation set are used for evaluating the models. The few shot prompt includes all 300 sentences and the assigned label in the conversion history of Gemini before classifying new excerpts.

Architecturally similar to ChatGPT, Gemini are a family of decoder-only transformer models that are developed by Google. The mid-tier version called Gemini Pro is found to perform in between GPT-3.5 and GPT-4 on a number of LLM evaluation benchmarks (Gemini Team, 2023).[13] Our results in Table 5 confirm this finding showing high performance of Gemini Pro closer to GPT-4 than GPT-3.5. The large context window of Gemini Pro, which spans 32k tokens, significantly more than the 4k token limit of GPT-3.5-0301, enables the inclusion of the 300 training sentences directly within the prompt history, acting as a "few shot" learner. This approach further boosts performance, making Gemini almost perform on GPT-4 levels. The fact that fine-tuned GPT-3.5, GPT-4 and Gemini can outperform 3.5 on a prompt that was engineered using GPT 3.5 is further evidence that model characteristics play a larger role than prompt engineering can.

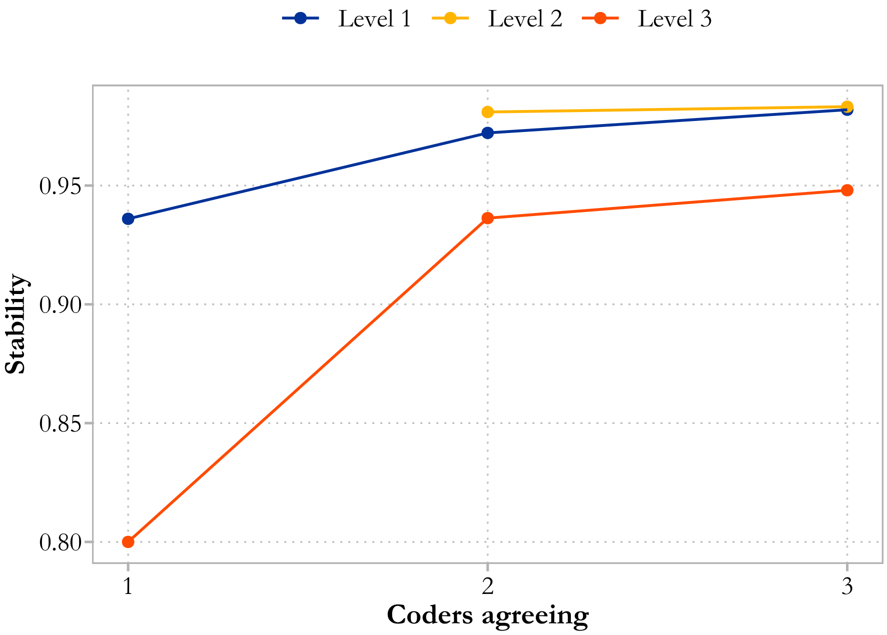
Finally, we can show that human and machine uncertainty about sentence classification correlate. For this, we test whether disagreement among human coders is related to ChatGPT's variation across runs. We document that ChatGPT's classification is more uncertain, meaning it varies more between different attempts across runs, for sentences that lack consensus among humans. Figure 6 plots the stability of ChatGPT's classification, measured as the share of sentences which ChatGPT classifies in its most frequent category across 25 runs, against the number of human coders who agree. Higher agreement among coders is associated with a more stable ChatGPT classification[14]. This variation could be used to construct an uncertainty measure. However, due

[13] Currently only Gemini Pro is available via the API. There are also Gemini Ultra, which is the most capable Gemini model, and Gemini Nano that is optimized for memory usage and computation speed.

[14] Unlike our main analysis, this test was run with temperature set to 0.25, which is still far below the

to the size of our dataset, running the entire dataset multiple times is not practical due to cost and token limit constraints.

FIGURE 6. Share of sentences classified with most frequent label by agreement of human coders



Note: Stability is measured as share of sentences that are classified identically when running ChatGPT-3.5-0301 25 times with temperature 0.25 and our final prompts

3.5. From raw data to constructing indices

To construct our full dataset, we run the final prompts on our pre-processed dataset consisting of 2,034,313 sentences. We randomly reshuffle the dataset to ensure that the sentences presented to ChatGPT inside the prompt are not subsequent sentences inside the speeches. We adopted this approach for three reasons (i) to mitigate the risk of ChatGPT categorizing sentences in relation to one another within a single speech, (ii) to ensure that accuracy observed in our random validation set is indicative of the whole dataset, and (iii) to decorrelate the classification of a speech with eventual fluctuations in model behavior over time. ChatGPT sometimes diverges from the specified output categories or formats in our prompts. We identified the most common deviations and categorized them accordingly when the classification was unambiguous. Sentences that were not successfully classified are tried a second time. After the two attempts,

default value of 1, as otherwise the variation is very small.

the number of sentences that remain unclassified is negligible across all three levels, with level 3 having the highest count of unclassified sentences at 1,433 instances, which represents only 0.07% of the dataset.

Our classified sample is tagged with the date of the speech and the central bank. We aggregate on the year-central bank level. Our indices of fiscal, monetary and financial dominance and coordination are turned into a proportional index ψ. We calculate the relative proportion of each of the categories $d \in D$ for all central bank $c \in C$ and year $t \in T$ combinations:

$$\psi_{c,t,d} = \frac{\sum_{j \in J} 1(\text{Classification}_j = d)}{\sum_{j \in J} 1(\text{Classification}_j \in D)}$$

with

$$\begin{aligned}
&D = \{\text{Monetary Dominance, Fiscal Dominance, Financial Dominannce,}\\
&\text{Monetary-fiscal coordination, Monetary-financial coordination}\}\\
&C = \{\text{European Central Bank, Bank of England,} \dots\}\\
&T = \{1997, \dots, 2023\}
\end{aligned}$$

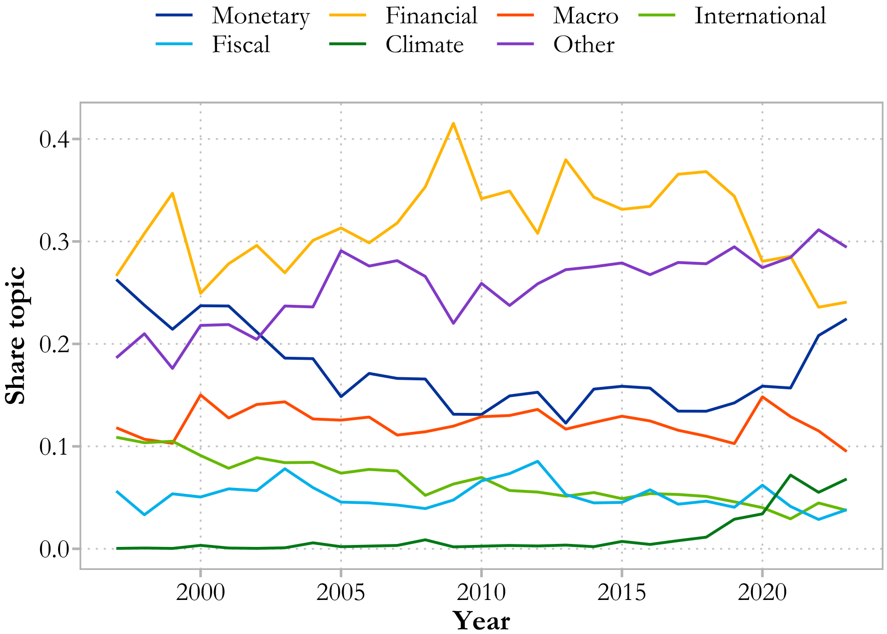
where J is the set of all sentences belonging to the central bank-year combination (c, t). The proportions together add up to 1 which creates desirable properties for further analysis and cross-central bank and time comparison for three reasons. First, it normalises the score for the increase in central bank communication over time. Second, it normalises the score taking into account the strong heterogeneity between central banks in their frequencies of communication. Third, it creates a relative measures of policy linkages, disregarding non-relevant communication. Different kind of policy linkages – dominance and coordination – are to be interpreted as a trade-off, namely more monetary dominance implies less fiscal and financial dominance and coordination. A relative index is thus an intuitive way to describe the linkage of monetary and other policies.

4. Descriptive variation of the indices

In the following, we first present the descriptive results of all three coding levels, including level 3 which provides us with indices of central bank policy-linkages across time and central banks. Second, we offer some first analytical insights from simple correlations of our indices with various political-economic variables, including (i) levels of inflation, (ii) economic development levels and political regimes as well as (iii) fiscal crises.

Figure 7 below shows the raw frequency of the number of sentences of our level 1 topic classification over time. Plotting and examining this frequency data allows for face validity of our produced classification. For example, given the role of central banks as counterparts of the financial sector - providers of liquidity, lenders of last resort and – in many – cases banking supervisors, central bankers refer to financial developments more often than any other policy area. Notably, the frequency of financial references picked up considerably after the Global Financial Crisis (GFC) and has only recently receded. Similarly, references to climate change have recently gained traction in central bank communication from 2016 onwards, in line with known trends.

FIGURE 7. Level 1 shares of the classification results of the topic model over time.



These topic frequencies differ strongly among sub-groups of countries (see Table 6). First, central bank officials of developing and emerging economies – following the classification of IMF (2023) – talk less about monetary topics, on average, and more about financial topics. This may partly be driven by earlier financial crises in developing countries, such as the 1997 Asian financial crisis and higher financial pressure in general. In contrast, central bank officials of advanced economies communicate more about monetary policy. Nevertheless, advanced economy central banks also make frequent reference to financial topics albeit less than developing and emerging countries. Second, similarly, we find that central banks in democratic countries talk considerably more about monetary topics and less about financial topics, compared to autocratic countries.[15] Third, one can examine the relationship with macroeconomic variables such as inflation (we discuss additional macro-economic variables in the Appendix). The frequency with which central bankers from countries with high levels of inflation refer to monetary policy or macroeconomics is somewhat lower than in central banks in low inflation countries. This can be partly due to central banks of high inflation countries having a range of other issues putting pressure on them being able to communicate about their core tasks.

TABLE 6. Distribution of topics

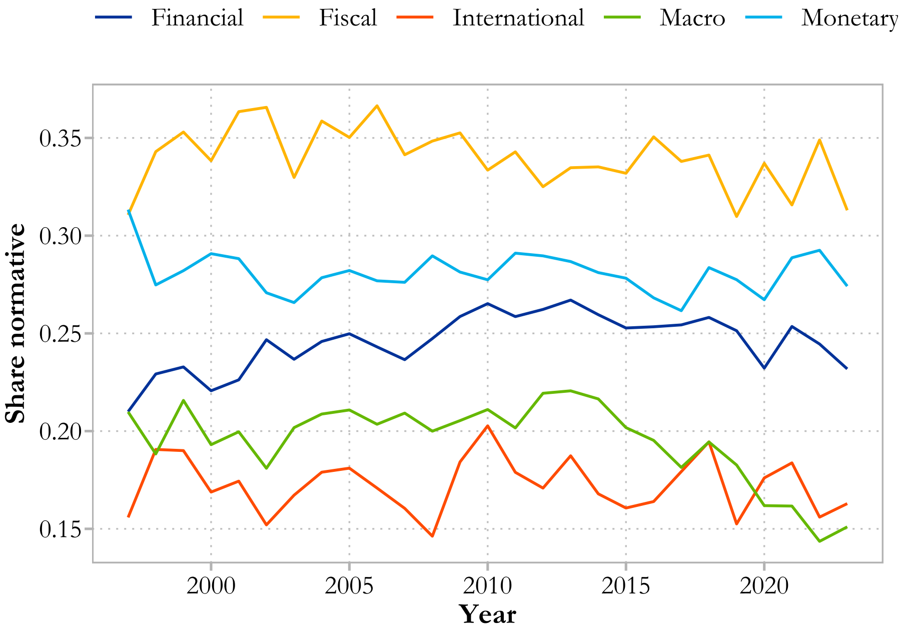
Country group	Monetary	Financial	Macro	International	Fiscal	Climate	Other
Advanced	20.1%	32.0%	13.5%	5.5%	5.1%	1.3%	22.3%
Emerging and Developing	12.5%	37.4%	10.7%	5.9%	4.0%	1.2%	28.2%
Democracy	19.4%	32.2%	13.4%	5.6%	5.1%	1.0%	23.1%
Autocracy	8.8%	42.2%	9.1%	5.8%	3.2%	2.0%	28.9%
High inflation	14.0%	35.4%	10.8%	5.7%	4.3%	1.0%	28.6%
Low inflation	17.5%	34.2%	13.4%	5.7%	4.6%	1.5%	23.1%

Note: Classification of advanced economies according to IMF (2023). Democracy indicator from the VDEM Dataset(Coppedge et al., 2023). High inflation countries are defined as countries with median inflation higher than the median inflation of the entire dataset from 1997-2023 (3.1%)

Figure 8 shows the level 2 classifications portrayed as the share of sentences classified as normative by topic. Central bank communication is most normative regarding fiscal policy. This may be due to the nature of the fiscal communication which will be uncovered in level 3. Communication on domestic and international macroeconomic developments is least normative which may reflect that when central bankers talk about macroeconomics they are often merely describing economic conditions.

[15] We follow Coppedge et al. (2023) whether a country is a democracy ("liberal and electoral") or autocracy ("closed and electoral").

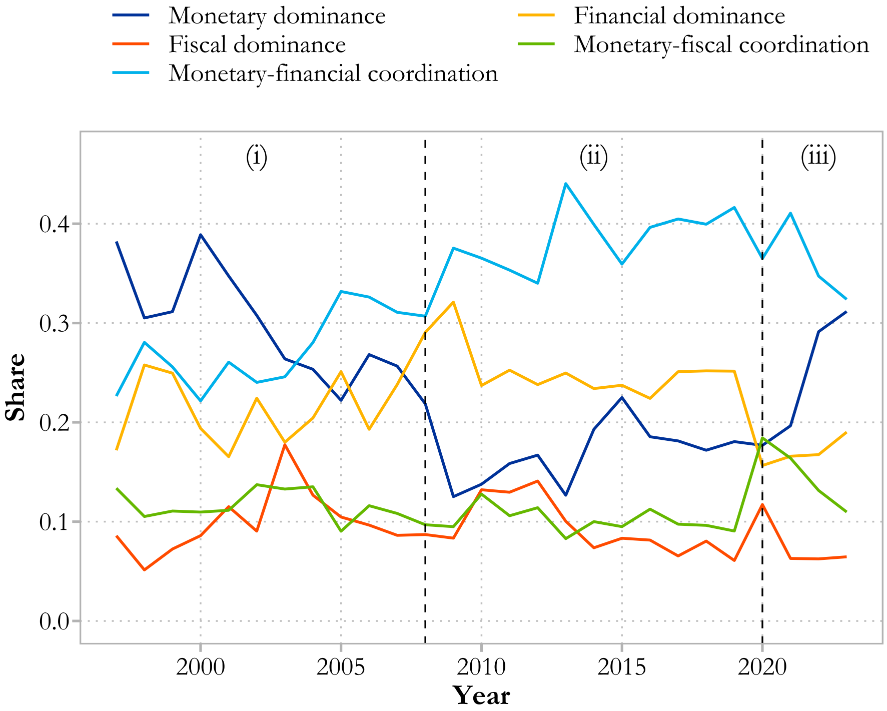
FIGURE 8. Level 2 classification results of the descriptive/normative classification.



Note: The Figure shows the share of normative sentences by topic over time. The topics 'other' and 'climate' are excluded. For 'climate' the share of normative sentences cannot be meaningfully calculated as communication on climate before 2016 was virtually zero.

Figure 9 below shows the classifications of level 3, which provides us with indices of linkages between monetary, fiscal and financial policy and their nature. Pooled across all central banks, our indices cover roughly three periods: (i) the pre-GFC period, (ii) the global financial crisis period and its aftermath, and (iii) the period from 2020 onward including the pandemic and the energy crisis. During these periods, one can observe three trends. First, prevalence of monetary dominance communication changed significantly over time. Namely, the pre-crisis period is characterised by a high degree of monetary dominance while following the GFC in 2007 there is marked drop in such rhetoric, which again strongly rebounds from 2020 onwards. Second, the chart also shows the indices for monetary-financial linkages. The post-GFC period shows a pickup of rhetoric that point both to monetary-financial coordination as well as financial dominance communication. Both receded, however, during the pandemic and energy crisis. Third, the chart shows communication on monetary-fiscal linkages which remains for both fiscal coordination as well as monetary-fiscal coordination relatively muted and stable over time.

FIGURE 9. Proportion of our policy-linkage indices pooled across all central banks over time of all sentences identified as concerning monetary, fiscal or financial.

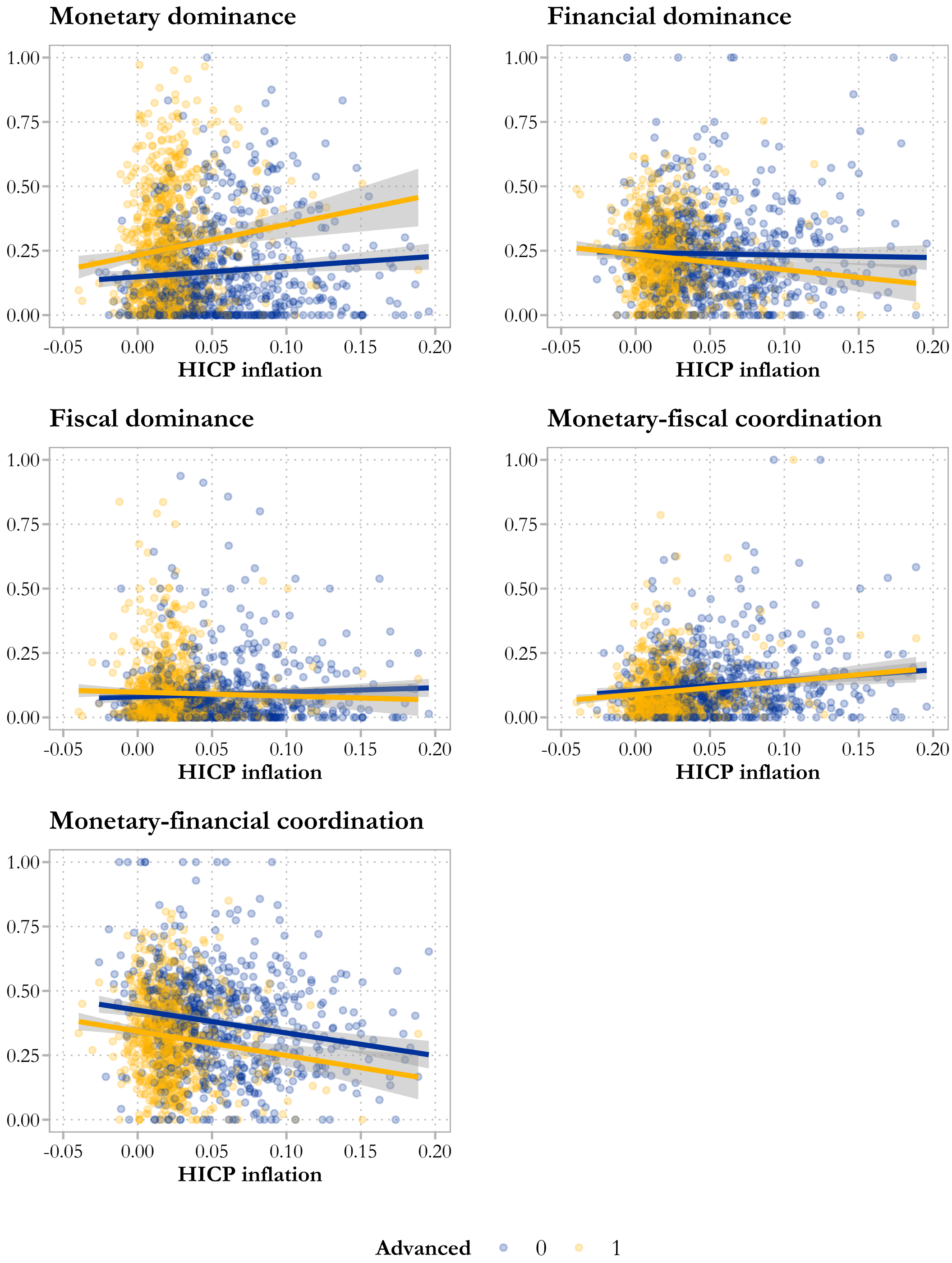


Note: The share shown is the relative index of dominance and cooperation as defined in section 3.5

To highlight possible applications of the indices in economic or political economy research we examine some basic correlations with (i) levels of inflation, (ii) economic development levels and political regimes and (iii) fiscal crises. Some first indicative results emerge that may motivate further research, three of which we present briefly in the following.

First, there is a positive relationship between the headline inflation rate and monetary dominance communication, particularly in advanced economies. Figure 10 shows this in the form of a scatterplot of HICP inflation and central bank-year observations of our monetary dominance indicator. This correlation may explain the observed variation of monetary dominance rhetoric over time. Namely, the post-GFC period was marked by low inflation and even deflationary pressures in many countries. Strong rhetoric on monetary dominance may have seemed out of the place at the time or even counter productive where central bank experienced an undershooting of their inflation target. Supporting this, there has been a marked pickup in inflation since the pandemic and energy crisis (the third period in our data), which has gone hand-in-hand with a significant revival of monetary dominance rhetoric. Thus, in times of high inflation, central bankers may resort to strong anti-inflationary rhetoric to bolster their credibility. The heterogeneity between central banks of advanced and developing economies may be explained by Figure 10 which shows that monetary dominance communication is more prevalent in democratic than in autocratic countries which is highly correlated with advanced and non-advanced economies, respectively. In this context, one could argue that given the more limited *de facto* independence of central banks in autocratic and less developed countries – there is less leeway for central bankers to use strong, anti-inflationary rhetoric in these countries. Additionally, Figure 10 also shows that higher inflation correlates with more monetary-fiscal coordination, both in advanced and developing economies. Thus, one could argue that in addition or instead of monetary dominance, central banks may also attempt to coordinate monetary with fiscal policy when inflationary pressures are higher.

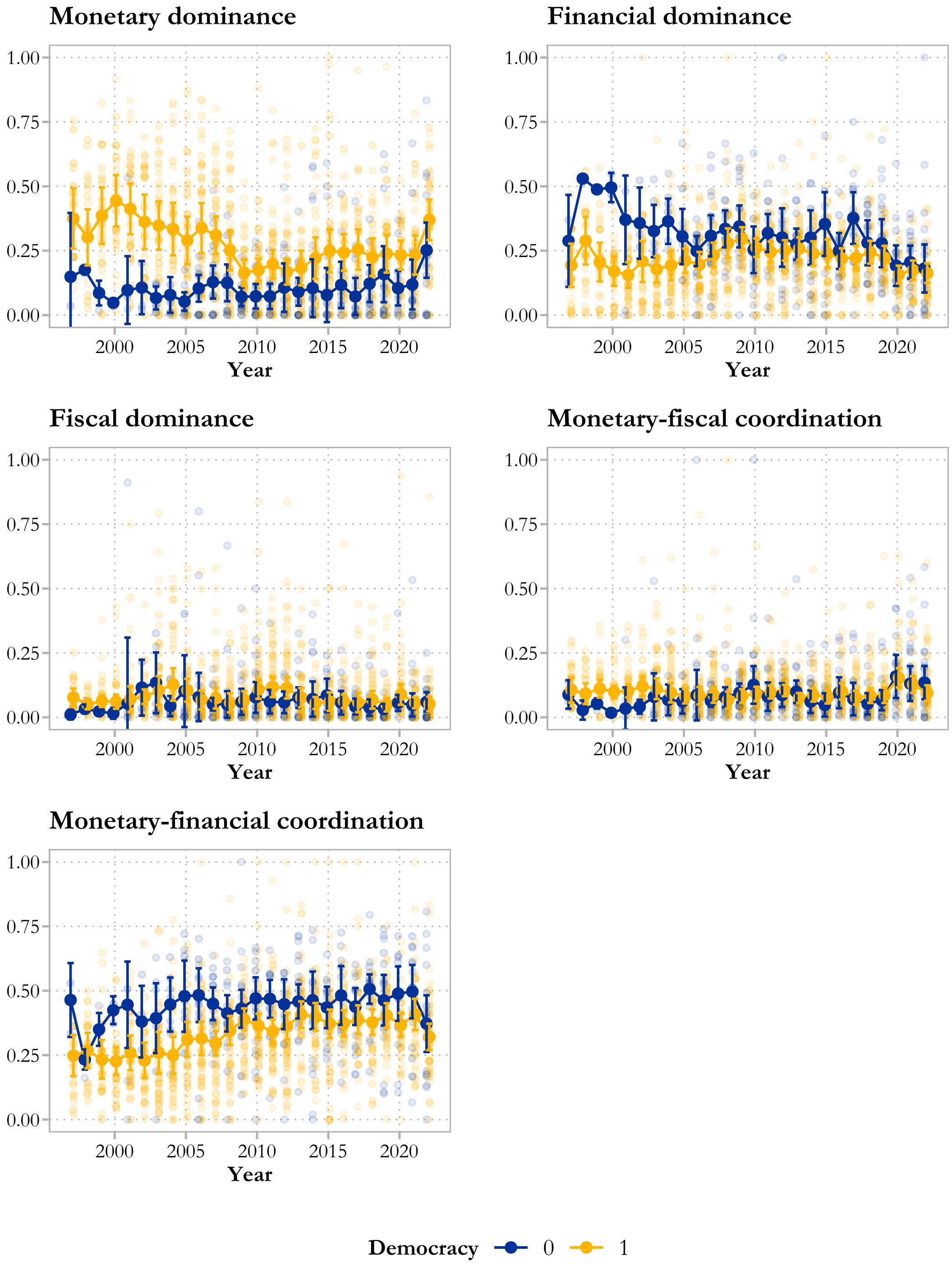
FIGURE 10. Scatterplots of the levels of inflation and their relationship with our policy-linkage indicators.



Note: Dots show central bank-year observations. The solid line indicates indicates a linear fit by advanced/emerging and developing economies. The shaded region around the regression line is the 95% confidence band.

Second, monetary-financial linkages play a larger role in autocratic and developing countries compared to in advanced and democratic countries. The scatterplots in Figure 11 for our various policy-linkage indicators clearly show this. This should not come as a surprise since central banks of autocratic, developing and emerging economies are often highly vulnerable to external financial crises, ranging from systemic banking, currency and sovereign debt crises which originate from the financial markets (see Valencia (2018). Moreover, developing countries with limited borrowing capabilities and small government sectors are constrained in their ability to react to financial stress. Namely, they often cannot make credible guarantees on bank liabilities or bail out defaulting financial institutions. In the literature, financial dominance has mainly been modelled formally (e.g., see Farhi and Tirole (2012)) following the idea that private borrowers and banks increase leverage because they anticipate that in a crisis the central bank will rescue them. In line with this, Gros and Shamsfakhr (2021) put forward multiple indicators to recognise financial dominance including indicators of excessive leverage or credit. Arguably, most of these indicators are more prevalent in autocratic and developing and emerging economies. This can also explain why the term "financial dominance" was first used by Fraga, Goldfajn, and Minella (2003) in the context of emerging market economies with inflation-targeting regimes. Nevertheless, financial dominance is also present, albeit to a lesser degree, in developed and democratic countries, especially after the GFC. For instance, Diessner and Lisi (2020) shows this for the case of the European Central Bank. Similar trends hold for monetary-financial coordination. Notable is also that these differences mainly are present before the financial crisis while there seems to be a slight convergence since the GFC. This can be explained by the fact that following the GFC, advanced economies saw themselves confronted with similar problems developing and autocratic countries had already experienced. One can see that this convergence is less strong for monetary-financial coordination, namely there is a continuously higher need to coordinate for non-democracies. This latter finding could also indicate that autocracies have to do more efforts in general to contain market pressures.

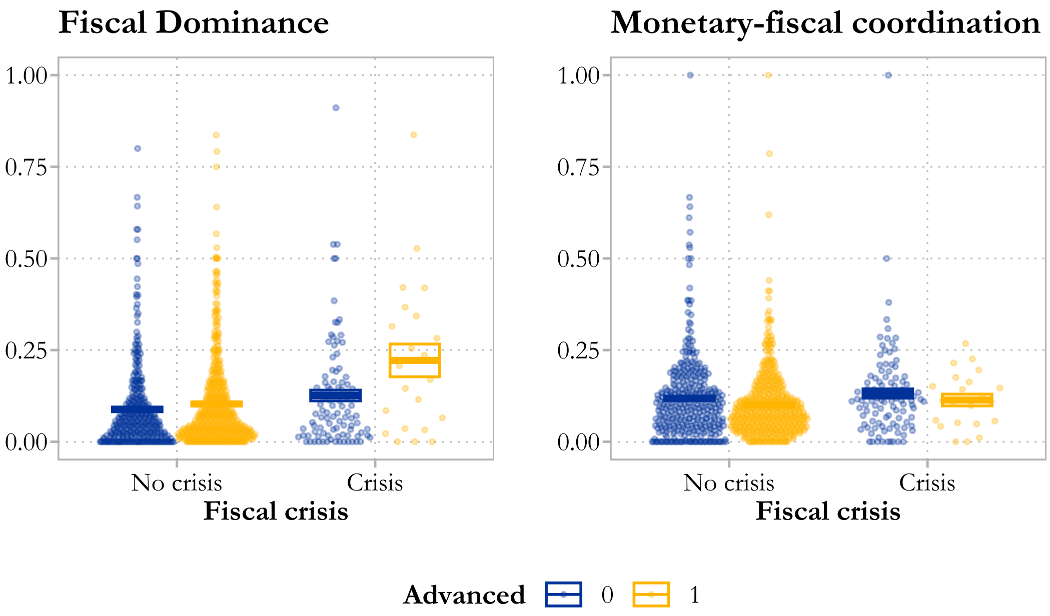
FIGURE 11. Scatterplots of the evolution our policy-linkage indicators in democracies and autocracies over time.



Note: Small dots show central bank-year observations. Observations with more speeches are more opaque. The solid line indicates the average of the category-average weighted by the number of speeches. Error bars indicate the 95% confidence interval of the weighted mean.

Third, fiscal crises have lead to an increase in fiscal dominance rhetoric, especially in central banks of advanced economies. This is shown in Figure 12 for which we have merged our dataset with the binary fiscal crisis indicator from Xu (2017). These trends are again unsurprising when one examines the literature. In the literature, there have been multiple conditions identified under which fiscal dominance is more likely to occur including, for instance, high net public liabilities, whether a fiscal authority has exhausted its fiscal capacity already and deliberate fiscal action of flooding the bond market to force central banks to undertake action (Mengus, Plantin, and Barthelemy, 2021). Therefore, communication regarding fiscal dominance can expected to be higher under these circumstances which is the case when fiscal crises events occur. Although fiscal dominance has been identified as a phenomenon in both advanced and non-advanced countries (e.g., Sabaté, Gadea, and Escario (2006) for an advanced country and see Makochekanwa (2008); Ersel and Özatay (2008) for two examples of non-advanced countries), in advanced countries fiscal crisis events are less frequent and, therefore, the reaction in communication can be expected to be stronger if it does occur. For instance, in the aftermath of the GFC, European countries were faced with severe tensions in the sovereign debt markets and instances of sovereign debt crisis which resulted in strong increases in fiscal communication. In contrast, Figure 12 shows that monetary-fiscal coordination does not show strong differences between crisis and non-crisis times and advanced and non-advanced (developing and emerging) countries. This can point towards fiscal-monetary interactions taking place both in normal and in crisis times and being less controversial in communication.

FIGURE 12. Violin plot of dominance and monetary-fiscal coordination during crisis and non-crisis times



Note: Fiscal crisis indicator from the IMF's fiscal crisis dataset (Xu, 2017). Each dot represents a central bank-year observation. Thick horizontal bars are estimates of the mean in each group. The box around the bars indicate the 95% confidence interval.

5. Conclusion

This paper has developed and (manually) validated novel textual indices using ChatGPT to examine central bank policy linkages with governments and financial markets expressed through monetary, financial and fiscal dominance and policy coordination from 1997 to mid-2023 for 118 central banks. Using these indicators, we then provided some descriptive statistics and correlations. Overall, this paper shows that Large Language Models, in particular ChatGPT, are useful for classification tasks. In contrast to existing tools (such as word counts), these tools allow to pursue complex classification tasks that are challenging even for human coders. We provide various results from our prompt engineering and reflect on efficiency choices which can be used as a stepping stone for similar classification tasks given the scarce number of use cases of Large Language Models in political economy and economics thus far. Beyond this, we also offer evidence that newer models can offer potential improvement in future use cases.

Given that our indicators are based on central bank communication, further research can examine the relationship between actual central bank policies and other empirical model approaches (e.g., DSGE models) which traditionally have been used to detect monetary and fiscal dominance regimes. Moreover, although this paper is mainly about measurement, we lay the groundwork for exploring relationships of the presented indices with various other variables. For instance, further research can examine how political variables specific to democratic and autocratic regimes, such as the impact of ideologically extreme governments, weak coalitions and military coups influence these indices. One can also examine the relationships with institutional variables, for instance, future research can tease out the relationships between changes in central bank independence indicators and the various indicators.

References

Afonso, António, and Priscilla Toffano. 2013. "Fiscal Regimes in the EU."

Ahmed, Rashad, Joshua Aizenman, and Yothin Jinjarak. 2021. "Inflation and Exchange Rate Targeting Challenges Under Fiscal Dominance." *Journal of Macroeconomics* 67: 103281.

Aruoba, Boragan, and Thomas Drechsel. 2022. "Identifying Monetary Policy Shocks: A Natural Language Approach."

Baerg, Nicole, and Will Lowe. 2020. "A textual Taylor rule: Estimating central bank preferences combining topic and scaling methods." *Political Science Research and Methods* 8 (1): 106–122.

Bennani, Hamza, and Matthias Neuenkirch. 2017. "The (home) bias of European central bankers: new evidence based on speeches." *Applied Economics* 49 (11): 1114–1131.

Bianchi, Francesco, and Cosmin Ilut. 2017. "Monetary/Fiscal policy mix and agents' beliefs." *Review of Economic Dynamics* 26: 113–139.

BIS. 2016. "Macroprudential Policy."

Blinder, Alan S., Michael Ehrmann, Jakob de Haan, and David-Jan Jansen. 2022. "Central Bank Communication with the General Public: Promise or False Hope?."

Brunnermeier, Markus K. 2015. *Financial Dominance*.: Banca d'Italia.

Chen, Xiaoshan, Eric M. Leeper, and Campbell Leith. 2022. "Strategic interactions in U.S. monetary and fiscal policies." *Quantitative Economics* 13 (2): 593–628. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1678.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M.Steven Fish Agnes Cornell, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Allen Hicken, Anna Lührmann, Seraphine F. Maerz, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Pamela Paxton, Daniel Pemstein, Johannes Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundtröm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2023. "'V-Dem Codebook v11' Varieties of Democracy (V-Dem) Project."

Cukierman, Alex, Steven B. Webb, and Bilin Neyapti. 1992. "Measuring the Independence of

Central Banks and Its Effect on Policy Outcomes." *The World Bank Economic Review* 6 (3): 353–398. Publisher: Oxford University Press.

Davig, Troy, and Eric M. Leeper. 2009. "Monetary-Fiscal Policy Interactions and Fiscal Stimulus."

Davig, Troy, and Eric M. Leeper. 2011. "Monetary–fiscal policy interactions and fiscal stimulus." *European Economic Review* 55 (2): 211–227.

Diessner, Sebastian, and Giulio Lisi. 2020. "Masters of the 'masters of the universe'? Monetary, fiscal and financial dominance in the Eurozone." *Socio-Economic Review* 18 (2): 315–335.

ECB. 2020. "The shadow of fiscal dominance: Misconceptions, perceptions and perspectives."Technical report.

Ersel, Hasan, and Fatih Özatay. 2008. "Fiscal Dominance and Inflation Targeting: Lessons from Turkey." *Emerging Markets Finance and Trade* 44 (6): 38–51.

Farhi, Emmanuel, and Jean Tirole. 2012. "Collective Moral Hazard, Maturity Mismatch, and Systemic Bailouts." *American Economic Review* 102 (1): 60–93.

Favero, Carlo A., and Tommaso Monacelli. 2005. "Fiscal Policy Rules and Regime (In)Stability: Evidence from the U.S.."

Ferrara, Federico Maria. 2020. "The battle of ideas on the euro crisis: evidence from ECB intermeeting speeches."

Fraga, Arminio, Ilan Goldfajn, and André Minella. 2003. "Inflation Targeting in Emerging Market Economies."

Gemini Team, Google. 2023. "Gemini: A Family of Highly Capable Multimodal Models."Technical report.

Gros, Daniel, and Farzaneh Shamsfakhr. 2021. "Financial Dominance: Not an Immediate Danger."

Haldane, Andy. 2020. "What has central bank independence ever done for us?."

Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. "Can ChatGPT Decipher Fedspeak?."

Hansen, Stephen, and Michael McMahon. 2016. "Shocking language: Understanding the macroeconomic effects of central bank communication." *Journal of International Economics* 99: S114–S133.

Hansson, Magnus. 2021. "Evolution of topics in central bank speech communication." arXiv:2109.10058 [econ, q-fin].

Heinemann, Friedrich, and Jan Kemper. 2021. "The ECB Under the Threat of Fiscal Dominance – The Individual Central Banker Dimension." *The Economists' Voice* 18 (1): 5–30.

Hinterlang, Natascha, and Josef Hollmayr. 2022. "Classification of monetary and fiscal dominance regimes using machine learning techniques." *Journal of Macroeconomics* 74: 103469.

IMF. 2023. "World Economic Outlook Database."

Laskar, Md Tahmid Rahman, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. "A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets." arXiv:2305.18486 [cs].

Leeper, Eric M. 1991. "Equilibria under 'active' and 'passive' monetary and fiscal policies." *Journal of Monetary Economics* 27 (1): 129–147.

Maddaloni, Angela, Caterina Mendicino, and Luc Laeven. 2022. "Monetary and macroprudential policies: Trade-offs and interactions."

Makochekanwa, Albert. 2008. "Impact of Budget Deficit on Inflation in Zimbabwe." *Economic Research Guardian*.

Marozzi, Armando. 2021. "Beware of Fiscal Signalling. The Effects of the ECB's Fiscal Communication in the Euro Area."

Mengus, Eric, Guillaume Plantin, and Jean Barthelemy. 2021. "Large public debts need not imply fiscal dominance."

Miranda-Agrippino, Silvia, and Hélène Rey. 2020. "U.S. Monetary Policy and the Global Financial Cycle." *The Review of Economic Studies* 87 (6): 2754–2776.

Moschella, Manuela, and Nicola Diodati. 2020. "Does politics drive conflict in central banks' committees? Lifting the veil on the European Central Bank consensus." *European Union Politics*.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. "Training language models to follow instructions with human feedback." arXiv:2203.02155 [cs].

O'Connor, Cliodhna, and Helene Joffe. 2020. "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines." *International Journal of Qualitative Methods* 19: 1609406919899220. Publisher: SAGE Publications Inc.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (85): 2825–2830.

Rogoff, Kenneth. 1985. "The Optimal Degree of Commitment to an Intermediate Monetary Target." *Quarterly Journal of Economics* 100: 1169–1189.

Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review* 94 (4): 1055–1084.

Sabaté, Marcela, Regina Escario, and Maria Dolores Gadea. 2015. "Fighting fiscal dominance. The case of Spain, 1874-1998." *European Review of Economic History* 19 (1): 23–43. Publisher: Oxford University Press.

Sabaté, Marcela, María Dolores Gadea, and Regina Escario. 2006. "Does fiscal policy influence monetary policy? The case of Spain, 1874–1935." *Explorations in Economic History* 43 (2): 309–331.

Sargent, Thomas J., and N Wallace. 1981. "Some Unpleasant Monetarist Arithmetic." *Quarterly Review* 5 (3).

Schelkle, Waltraud. 2023. "Monetary re-insurance of fiscal states in Europe." *SM*: 29–52. Publisher: Societ editrice il Mulino Section: 1/2023.

Schonhardt-Bailey, Cheryl. 2013. *Deliberating American Monetary Policy.*: MIT press.

Stiglitz, Joseph E. 2017. "Where Modern Macroeconomics Went Wrong."

Strasser, Georg, Gaetano Gaballo, Peter Hoffmann, and Michael Ehrmann. 2019. "Signalling a future path of interest rates: The international evidence on forward guidance."

Swanson, Eric T. 2021. "Measuring the effects of federal reserve forward guidance and asset purchases on financial markets." *Journal of Monetary Economics* 118: 32–53.

Valencia, Luc Laeven, Fabian. 2018. "Systemic Banking Crises Revisited."

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv:1706.03762 [cs] version: 5.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv:2201.11903 [cs].

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. "HuggingFace's Transformers: State-of-the-art Natural Language Processing." arXiv:1910.03771 [cs].

Xu, Tigran Poghosyan Juan Farah-Yacoub Yizhi Kerstin Gerling, Paulo A. Medas. 2017. "Fiscal Crises."

Appendix A. Replication files on Github

We provide a repository with replication codes on Github. It contains all the codes to pre-process the dataset, run ChatGPT on two million sentences, and finally produce our indices and aggregated results. Moreover, we provide our manually classified validation sample
`inputdata/validation_sample_all.xlsx` and the codes to conduct prompt engineering experiments, fine-tune GPT-3.5, and assess the classification quality of various ChatGPT models and Gemini Pro against this validation set. The following contains the descriptions of the included files and their function. The same instructions can be found inside the README.md.

We share a yearly aggregation of our indices of dominance and coordination `dominance_coordination_dataset.csv`. This file is sufficient to produce all charts inside the Appendix and main part of the paper. Importantly, we don't include any speeches or sentence-level results. The output files are more than a gigabyte in size and too large for this repository. To rerun the full analysis, the speech data would need to be scraped with the python code here. We do, however, provide the sentence-level classification of our prompt engineering results, validation exercise, and model comparisons. These are stored as Pandas DataFrames in `.pkl` format inside the `outputdata` folder.

A.1. Instructions to run codes

- To rerun any of our analyses, an API key for ChatGPT and/or Gemini needs to be set inside the `llm_functions.py` file. Also note that these LLMs, even at a temperature set to zero, are non-deterministic. Exact results vary with each run, although with ChatGPT, usually 97%-99% of sentences are identically classified across two runs. In addition, changes to the model on OpenAI's/Google's side can impact results.
- To run R codes, the working directory should be set to the root of the project.
- Python codes expect to be run from the folder they are in.
- Validation, prompt engineering, and model comparison codes are self-contained and can be run with the inputs provided inside this repository, provided that an API key is set.

A.2. Included files

The `codes` folder contains the following files:

- 0_text_preprocessing.py: This file runs the preprocessing steps described in the Appendix.
- 1_chat_gpt_main_analysis.py: This code consists of the code required to run the full dataset. It requires the output produced by 0_text_preprocessing.py.
- 2_validation_and_robustness.py: This file contains the code for the robustness checks, prompt engineering results, and different ChatGPT versions. It requires only our validation set as input validation_sample_all.xlsx.
- 3_fine_tuning_and_few_shot.py: This file constructs a training dataset from our validation set, trains a fine-tuned GPT 3.5 model, and evaluates it with the remaining sample. Moreover, it contains code to run Gemini Pro using (i) the same prompts as ChatGPT and (ii) a few-shot prompting strategy.
- llm_functions.py: Functions that are shared by the python codes are in this file. Most notably, it contains the function that takes a dataframe as input and calls either the Gemini or ChatGPT API with our prompt design. This function allows for parallel API queries to maximize rate limits.
- merge_datasets.R: This R code calculates our relative indicator of dominance and coordination. It requires the outputs saved by 1_chat_gpt_main_analysis.py. It also sketches how our shared dataset dominance_coordination_dataset.csv is produced (without including the third-party data sources).
- run_all_charts.R: Produces all of the charts.

A.3. Replication of Charts

All our charts can be replicated with the R codes inside the codes/figures folder. Run run_all_charts.R to produce all charts. The R files read from the ChatGPT results provided inside the outputdata and the yearly aggregation of the full dataset dominance_coordination_dataset.csv. No access to ChatGPT is required to produce the charts. These are the files to produce the charts:

- bin_scatter.R
- correlation.R
- crisis.R
- levels_over_time.R
- sentence_count_charts.R
- stability.R
- temperature_charts.R

Common functions and settings to change the size of the charts are inside functions_and_settings.R.

A.4. Prompts

The instructions part of our prompts are stored in the prompts folder. The sentences/excerpts are automatically appended to the prompt. We use a .yaml format to store the prompts. Our final instructions for level 1, level 2, and level 3 are in the l1, l2, l3 subfolders. To change the prompts either modify the prompt file or modify the python code to load a different prompt.

Appendix B. Database Construction

This dataset of central bank official speeches is scraped from the Bank of International Settlement website. In mid-2023, we downloaded a total of 18,081 speeches, press conferences, interviews and lectures issued by 118 institutions, mainly central banks, over the period from 1997 to 2023. See Table A1 for the contained central banks, the number of speeches scraped and the first and last year speech obtained per institution. The speeches are downloaded in PDF format and converted to text files. The PDF format is not entirely standardised, which can result in the core text being interspersed with page numbers, footnotes, tables, charts, and literature references, thereby disrupting the flow of the main text. To clean the text we apply several pre-processing steps (see the next section B.1).

TABLE A1. List of central banks, time periods and number of speeches

Central Bank	First Year	Last Year	Speeches Count
Bank of France	1997	2023	377
Bank of Finland	1997	2023	173
Sveriges Riksbank	1997	2023	481
Reserve Bank of India	1997	2023	865
People's Bank of China	1997	2023	142
Bank of Italy	1997	2023	370
European Monetary Institute	1997	1997	2
South African Reserve Bank	1997	2023	393
Deutsche Bundesbank	1997	2023	802
Federal Reserve Bank of New York	1997	2023	411
Reserve Bank of Australia	1997	2023	528
Reserve Bank of New Zealand	1997	2023	190

TABLE A1. List of central banks, time periods and number of speeches *(continued)*

Central Bank	First Year	Last Year	Speeches Count
Czech National Bank	1997	2023	55
De Nederlandsche Bank	1997	2023	205
Bank of Canada	1997	2023	552
Banque Nationale Suisse	1997	1997	3
Central Bank of the Republic of Turkey	1997	2022	99
Hong Kong Monetary Authority	1997	2023	249
Bank for International Settlements	1997	2015	17
Federal Reserve Bank of Kansas City	1997	2011	19
Bank of Greece	1997	2023	157
Swiss National Bank	1997	2023	383
International Monetary Fund	1997	2014	4
Central Bank of Ireland	1998	2023	318
Central Bank of Iceland	1998	2023	90
Central Bank of the Republic of Austria	1998	2023	81
European Central Bank	1998	2023	2377
Bank of Korea	1998	2023	89
Central Bank of Norway	1999	2023	284
Central Bank of Brazil	1999	2015	11
Bank Indonesia	1999	2023	63
Monetary Authority of Singapore	1999	2023	290
National Bank of Belgium	1999	2022	38
Bank of Latvia	1999	2018	11
Bank of Namibia	1999	2022	32
Bank of Thailand	2000	2023	221
Croatian National Bank	2000	2023	10
Federal Reserve Bank of San Francisco	2000	2012	3
National Bank of North Macedonia	2001	2023	92
Bank of Israel	2001	2022	105
Central Bank of Bosnia and Herzegovina	2001	2020	15
Central Bank of Trinidad and Tobago	2001	2018	101
Central Bank of Chile	2001	2023	125
Bank of Estonia	2001	2023	21
State Bank of Pakistan	2001	2022	132
Central Bank of Malta	2001	2022	58
Bank of Poland	2002	2007	20
Central Bank of Malaysia	2002	2023	487
National Bank of Denmark	2002	2023	101
Bank of Spain	2002	2023	329
Central Bank of Luxembourg	2002	2015	38

TABLE A1. List of central banks, time periods and number of speeches *(continued)*

Central Bank	First Year	Last Year	Speeches Count
Bank of Mauritius	2002	2023	160
Bank of Zambia	2003	2023	156
Bank of Botswana	2003	2022	46
Bank of Papua New Guinea	2003	2022	60
National Bank of Slovakia	2003	2019	5
Bank of Mexico	2003	2023	98
Central Bank of Nigeria	2003	2023	35
Reserve Bank of Malawi	2003	2017	25
Bank of Portugal	2003	2021	60
Eastern Caribbean Central Bank	2003	2019	17
Central Bank of the Bahamas	2004	2019	13
Bank of Albania	2004	2023	290
Central Bank of Barbados	2004	2023	91
Bank of Sierra Leone	2004	2011	12
Central Bank of Sri Lanka	2004	2019	67
Central Bank of the Philippines	2004	2023	521
Central Bank of Argentina	2004	2021	34
Reserve Bank of Fiji	2004	2021	132
National Bank of Serbia	2004	2023	114
Bank of Jamaica	2005	2023	25
National Bank of Romania	2005	2023	71
Bank of Uganda	2005	2019	152
Bank of Ghana	2005	2022	56
Bulgarian National Bank	2006	2023	44
Saudi Arabian Monetary Agency	2006	2014	28
Central Bank of Hungary	2006	2022	11
Central Bank of the United Arab Emirates	2006	2019	10
Bank of Mozambique	2006	2010	5
Reserve Bank of Vanuatu	2006	2013	2
Central Bank of Solomon Islands	2006	2019	18
Monetary Authority of Macao	2007	2014	27
Central Bank of Bahrain	2007	2017	45
Central Bank of Kenya	2007	2023	182
Central Bank of Colombia	2007	2009	3
Central Bank of Samoa	2007	2014	6
Bank of Algeria	2009	2022	8
National Bank of Cambodia	2009	2022	3
Central Bank of Aruba	2009	2009	1
Federal Reserve Bank of Boston	2010	2011	5

TABLE A1. List of central banks, time periods and number of speeches *(continued)*

Central Bank	First Year	Last Year	Speeches Count
Central Bank of Jordan	2010	2010	1
Central Bank of Bolivia	2010	2010	1
Central Bank of Cyprus	2011	2023	17
Federal Reserve Bank of Philadelphia	2011	2015	32
Federal Reserve Bank of Dallas	2011	2015	19
Central Bank of Belize	2011	2011	1
Central Bank of Curaçao and Sint Maarten	2011	2022	14
Federal Reserve Bank of Minneapolis	2011	2014	23
Federal Reserve Bank of Chicago	2011	2014	10
Bank of Guatemala	2012	2012	1
Central Bank of Uruguay	2012	2012	1
Bank of Guyana	2012	2012	2
Central Bank of Nepal	2013	2019	14
Bank of Tanzania	2013	2013	1
Bank of Slovenia	2014	2023	14
Bank of Russia	2014	2022	43
Federal Reserve Bank of Atlanta	2015	2015	2
Federal Reserve Bank of Richmond	2015	2015	2
Bank of Lithuania	2015	2022	34
Central Bank of Kuwait	2017	2022	5
Central Bank of Seychelles	2017	2023	23
Maldives Monetary Authority	2018	2021	5
Central Bank of the Republic of Kosovo	2018	2023	28
National Bank of Ukraine	2019	2023	33
Bank of Morocco	2019	2021	6

B.1. Pre-processing steps

All the text pre-processing steps are undertaken in Python and available in the replication files (see `text_processing.py` in the codes folder). The pre-processing steps follow the following steps:

a. **Regular expressions**
   We examined the extracted text for recurring patterns of:
   - Page numbers
   - Page headers
   - New page characters
   - Footnotes
   - URLs
   - Subsequent whitespace characters

   , which we remove using appropriate regular expressions.

b. **Conversion to sentence level**
   We convert the entire corpus to the sentence level using the *Punkt* sentence tokenizer from the NLTK python package.[16] We also tried the sentence extraction from spacy's *en_core_web_lg* model, which we found to produce similar results while being much slower.

c. **Sentence level heuristics**
   After segmenting the corpus into individual sentences, the dataset still contains entries that do not constitute genuine sentences of the primary text. Instead, these entries include tables, annotations, or binary data erroneously recognized as text during the PDF conversion process. To address this issue, we implement conservative rules aimed at filtering out clearly irrelevant entries:
   - Remove sentences with less than 2/3 ASCII characters
   - Remove sentences that consists of less than 6 tokens or more than 200 tokens
   - Remove sentences with less than 20 characters

After pre-processing, we obtain a dataset that consists of 2,034,313 sentences. A small share of sentences do not contain relevant text, i.e., some chart annotations, references etc. remain. We are reluctant to more aggressively delete sentences as ChatGPT will identify irrelevant sentences anyways. After pre-processing, we run the three levels of ChatGPT classification, construct our relative indicator and aggregate on the central bank-year level as described in section **??** of the main text.

[16] https://www.nltk.org/

B.2. Variables in the Database

Our shared database contains our index of the relative shares of monetary-dominance, fiscal-dominance, monetary-fiscal coordination, monetary-financial coordination and financial dominance as well as additional country-year level data on macroeconomic and political indicators. Table A2 describes the variables and where applicable their data sources will be provided.

TABLE A2. Variables in the database

Variable name	Description
Metadata	
central_bank	Name of the central bank
country	The ISO 3166-1 3 letter country code of the central bank's country
year	The year of the observation
currency_code	The ISO 4217 3 letter currency code of the central bank's currency
advanced	Takes value 1 if the country is a advanced economy according to the IMF classification of advanced, emerging and developing countries.
number_of_speeches	The number of speeches that were classified by the central bank in a given year
number_of_sentences	The total number of sentences that were classified by the central bank in a given year
Level 1	
topic_monetary	Share of sentences that address monetary issues (e.g., inflation, price stability, primary mandate, interest rate)
topic_fiscal	Share of sentences that address fiscal issues (e.g., sovereign debt, budget balance, fiscal governance, taxes, pensions)
topic_financial	Share of sentences that address financial issues (e.g., banking supervision, financial instability, credit risks)
topic_climate	Share of sentences that address climate issues (e.g., environmental issues, CO2, climate change, sustainable development goals)
topic_macro	Share of sentences that address macroeconomic issues (e.g., GDP, economic growth, unemployment, productivity, economic outlook)
topic_international	Share of sentences that address international economics issues (e.g., trade, exchange rates, capital mobility, tariffs)
topic_other	Share of sentences that address topics of monetary, fiscal, financial, climate, macro or international economics.
Level 2	
share_normative	Share of sentences that contain a value judgement.
share_descriptive	Share of sentences that are descriptive and do not provide a value judgement.
normative_<topic>	For each of the topics (see Level 1) the dataset contains a variables which describes the share of sentences belonging to the topic that are classified as normative

TABLE A2. Variables in the database *(continued)*

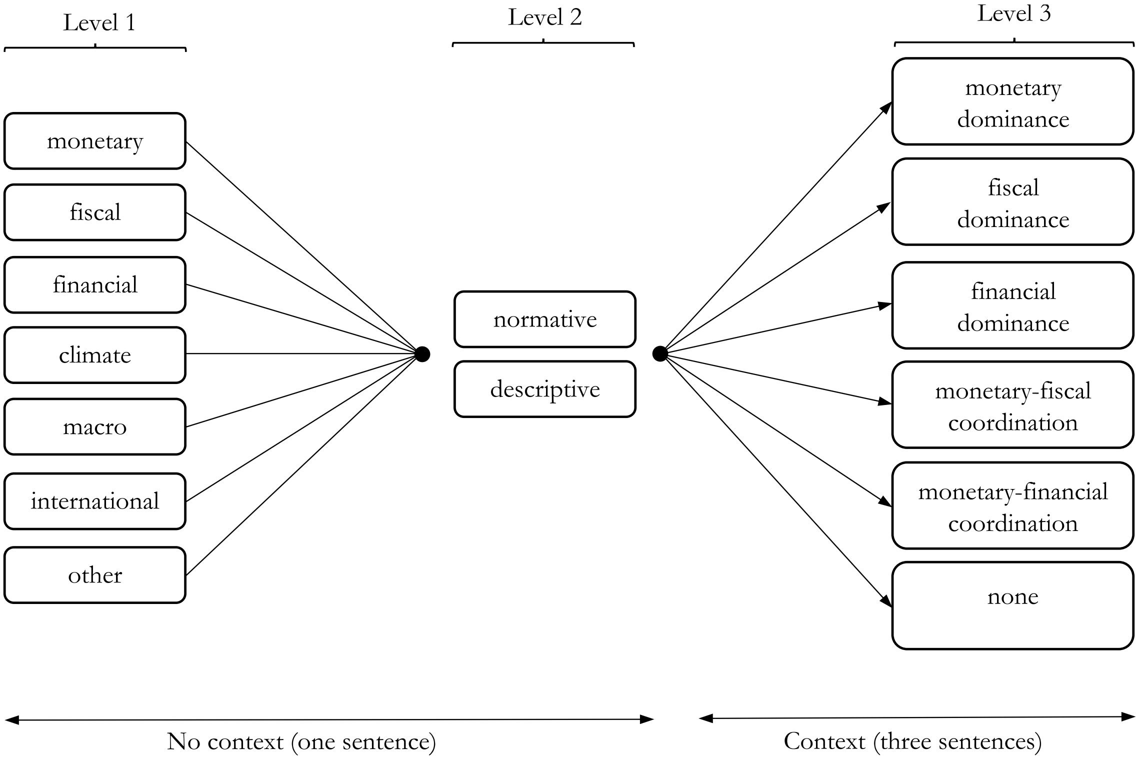
Variable name	Description
Level 3	
monetary_dominance	Share of excerpts that classify as "monetary dominance" ,i.e., if the excerpt clearly and explicitly says that the central bank subordinates fiscal or financial policies to the central bank's monetary policy objective of price stability
fiscal_dominance	Share of excerpts that classify as "fiscal dominance" ,i.e., if the excerpt clearly and explicitly says that the central bank subordinates itself to fiscal authorities, that is, where the speaker says that monetary policy is primarily driven by fiscal considerations rather than maintaining price stability
financial_dominance	Share of excerpts that classify as "financial dominance" ,i.e., if the excerpt clearly and explicitly says that the central bank subordinates to financial markets or the financial regulation authorities, that is, where the speaker says that monetary policy is primarily driven by financial stability considerations rather than maintaining price stability
monetary_financial_ coordination	Share of excerpts that classify as "monetary-financial coordination" ,i.e., if the excerpt suggests that the central bank and financial regulators should cooperate, this is, where the speaker says that monetary policy and financial regulation are best coordinated to achieve the right policy mix
monetary_fiscal_ coordination	Share of excerpts that classify as "monetary-fiscal coordination" ,i.e., if the excerpt suggests that fiscal authorities and the central bank should cooperate, this is, where the speaker says that monetary and fiscal policy are best coordinated to achieve the right policy mix
none	Share of excerpts that classify as "none", i.e., if there is no reference to monetary, financial or fiscal developments or if the excerpt describes monetary, financial or fiscal developments, that is, if the speaker does not make any normative reference to monetary, financial or fiscal policy.
Macroeconomic variables	
inflation	HICP inflation, IMF WEO variable code: PCPI
gdp_real_ppp_capita	GDP per capita measured in purchasing power parity (PPP) USD, IMF WEO variable code: NGDPRPPPPC
spread	Government bond spread measured as difference of 10Y government bond yield of the country to the German government bond yield. Data from Bloomberg, and in some cases where Bloomberg does not contain the relevant bond yield Refinitiv/Datastream
Political Variables	
democracy_ind	Dummy variable based on the v2x_regime variable from the VDEM dataset indicating a democracy. Democracy (1) includes electoral and liberal democracy; autocracy (0) includes closed and electoral autocracy
polarization_ind	Dummy variable based on the variable v2cacamps_mean from the VDEM dataset indicating high polarization, whereby 0 and 1 map to low polarization (0) and 2 and 3 map to high polarization (1).

Appendix C. Validation

C.1. Graphical coding overview

The coding of dominance and coordination is conducted in three steps. First, we classify the topic of the sentence as either ’monetary’, ’fiscal’, ’financial’, ’climate’, ’macro’, ’international’, or ’other’. The second step, level 2, determines whether the sentences is normative or descriptive. Level 3 then categorizes sentences as a form of dominance or coordination, also entailing the most frequent option of none. For level 3, the sentence before and after each sentence is added as context. The main text provides more details on why certain choices were made regarding the classification. See Figure A1 below for an overview of the coding scheme.

FIGURE A1. three-level coding classification overview.



C.2. Ambiguous sentence coding guidelines

We recognize that there are sentences for which the coding can be ambiguous. Therefore, these are some extra guidelines for possibly ambiguous cases.

a. Level 1:

- References to digital are classified as ‘other’
- Topics that refer to payment systems are classified as ’other’
- When a sentence discusses both monetary and fiscal topics, or both monetary and financial topics, it should be categorized under fiscal or financial, rather than under monetary.
- References to exchange rates should be classified as ’international’
- References to forecast and general developments in the economy should be classified as ’macro’

b. Level 2:
   - Classify as normative when words such as “should” or “our opinion” indicate a value judgement or a course of action.
   - Also implicit value judgements highlighting the importance of a certain topic can be classified as normative.
   - Any sentences that are a neutral recital of facts should be classified as descriptive.

c. Level 3:
   - If the excerpt alone does not provide sufficient context to identify the type of dominance or coordination, or if the context is assumed rather than explicitly provided, it should only be classified as a form of dominance or coordination if the context is widely acknowledged as common knowledge.
   - To differentiate between dominance and coordination, one should examine whether there are signs of a hierarchical relationship or an emphasis on interactions without a power hierarchy.

C.3. Classification Examples

The following Tables A3, A4, A5 provide examples of sentences from our validation set and their classification

TABLE A3. Level 1 classification examples

Classification	Example and explanation
Fiscal	Some countries, such as the United States and Ireland, also implemented extensive and costly government measures to support the financial sector. (Retrieved from: The Sveriges Riksbank, 18-10-2011). *Explanation:* the sentences mentions both fiscal and financial but the focus is on governments taking action.

TABLE A3. Classification examples *(continued)*

Classification	Example and explanation
Monetary	Among other things, this outcome complicates our ability to assess the present stance of monetary policy. (Retrieved from the Federal Reserve Bank of New York, 31-10-2006.) *Explanation:* The sentence describes an implication for their monetary policy.
Financial	Recent events have demonstrated the important role that banks play as liquidity providers and the potential for broader market turbulence when banks have difficulty performing this role. (Retrieved from the Board of Governors of the Federal Reserve System, 07-03-2008) *Explanation:* This sentence mentions how the role of bank's as liquidity providers is important for the economy.
Climate	In the longer run, the only way to address the climate crisis and to safeguard Europe's energy security is by accelerating the energy transition. (Retrieved from the Bundesbank, 15-11-2022). *Explanation:* The sentence mentions the importance of addressing the climate crisis.
International	However, experience shows that it is always correct to allow the exchange rate to change based on market forces. (Retrieved from the Bank of Israel, 23-04-2013). *Explanation:* Sentences referring to exchange rate are classified as international.
Macro	It does seem likely that productivity calculated for the entire economy using GDP data weakened in the second quarter. (Retrieved from the Board of Governors of the Federal Reserve System, 24-07-1998) *Explanation:* The sentence provides information on the macroeconomic outlook.
Other	The ECB's Legal Committee is a genuine example of the identity of the Eurosystem and the ESCB., (Retrieved from the European Central Bank, 24-01-2019) *Explanation:* The sentence talks about a legal committee which cannot be linked to any of the other categories thus by default will be 'other'.

TABLE A4. Level 2 classification examples

Classification	Example and explanation
Normative	Economic development calls for clear and predictable rules, and institutions that assure they will be enforced. (Retrieved from: The Bank of Mexico, 28-04-2003). *Explanation:* The sentence clearly entails a value judgement since the bank argues that rules enforced by institutions are necessary.
Descriptive	Such a large increase in income has only happened on a few occasions over the past decades. (Retrieved from the Sveriges Riksbank, 22-05-2002.) *Explanation:* The sentence relates macroeconomic conditions to the previous periods.

TABLE A5. Level 3 classification examples

Classification	Example
Monetary Dominance	"Furthermore, monetary policy implementation in line with the market efficiency principle would need to remain without prejudice to our primary mandate of safeguarding price stability." (Retrieved from: The European Central Bank, 14-06-2021). *Explanation:* The topic concerns a monetary topic and they emphasize their primary mandate of price stability being above other priorities. Therefore, this sentence can be classified as monetary dominance.
Fiscal dominance	"Moreover, although most of the resources administered by the BIS are invested in financial assets of top quality at international level and their exposure to the various risks are managed conservatively, a greater portion of such funds could be spend toward the direct purchase of debt denominated in local currencies of emerging countries or to the use of them as collateral of certain bond issuance of countries with limited depth of their financing markets in local currency." (Retrieved from the Central Bank of Argentina, 09-07-2008.) *Explanation:* This sentence refers to funds being spend towards the direct purchase of debt (=monetary financing) instead of considering pure price stability considerations, thus we consider this sentence to be fiscal dominance.
Financial dominance	"It is thus significant that our flexible and abundant provision of liquidity contained market participants' concerns over liquidity financing." (Retrieved from the Bank of Japan, 04-07-2002) *Explanation:* This sentence states that monetary policy is accommodating financial markets by providing liquidity, thus showing that financial markets are a consideration for the bank in conducting their monetary policy.

TABLE A5. Level 3 classification examples *(continued)*

Classification	Example
Monetary-fiscal coordination	"Since restarting our strategy review, we have introduced a new work stream on monetary-fiscal interactions precisely to address such questions." (Retrieved from the European Central Bank, 30-09-2020). *Explanation:* This sentence refers to the monetary-fiscal interactions which is a key policy in the monetary-fiscal coordination.
Monetary-financial coordination	"If market participants are willing to continue to work together, then we can safely achieve the transitions needed to create a better and more robust system that will help to ensure our ongoing financial stability." (Retrieved from the Board of Governors of the Federal Reserve System, 07-11-2017). *Explanation:* This sentence shows that the bank wants coordinate with market participants to ensure financial stability.

C.4. Coder reliability scores

The random sample of 1000 sentences was coded by the 3 authors who each have at least a MSc degree in Economics/Political Economy. Before starting the coding process, the authors went through the codebook and the examples provided above together, which have in turn been developed inductively based on another random sample of 100 sentences. All three human coders subsequently independently coded the same first 400 sentences. Table A6 shows a matrix of the agreement between the human coders (L,M,S) and ChatGPT (C). The complete dataset was extended to 1000 sentences by each author coding an additional 200 sentences.

We calculate Krippendorff's alpha that measures how well different coders agree when coding data at different levels of measurement. We find alphas of 0.81, 0.56 and 0.62 for level 1, level 2 and level 3 respectively among the human coders. In addition, we report Cohen's Kappa. Table A7 provides coder reliability scores with ChatGPT added as forth coder and the average reliability score of all possible combinations of replacing one human coder with ChatGPT.

C.5. Validation metrics

In the following we calculate validation metrics of our ChatGPT classifier that we compare against the "ground truth" of our human classification (see Table 3 of the main text). We differentiate between the full sample of 1000 sentences and a agreement sample, which consists of the sentences that all 3 human coders have classified identical. We rely mostly on F1 scores to assess our models. The F1 score is calculated as the

TABLE A6. Coder overlap matrices

A. Level 1

	L	M	S	C
L	1.00	0.87	0.88	0.85
M	0.87	1.00	0.81	0.77
S	0.88	0.81	1.00	0.78
C	0.85	0.77	0.78	1.00

B. Level 2

	L	M	S	C
L	1.00	0.81	0.88	0.85
M	0.81	1.00	0.78	0.73
S	0.88	0.78	1.00	0.76
C	0.85	0.73	0.76	1.00

C. Level 3

	L	M	S	C
L	1.00	0.84	0.87	0.62
M	0.84	1.00	0.82	0.60
S	0.87	0.82	1.00	0.61
C	0.62	0.60	0.61	1.00

Note: Cells of the matrix show the share of classifications that agree between the pair of coders.

TABLE A7. Coder reliability scores

	Level 1	Level 2	Level 3
A. Krippendorf's α			
Humans	0.81	0.56	0.62
One human replaced by ChatGPT	0.76	0.48	0.36
Humans + ChatGPT	0.77	0.50	0.42
B. Cohen's κ			
Humans	0.81	0.56	0.62
One human replaced by ChatGPT	0.76	0.48	0.38
Humans + ChatGPT	0.78	0.50	0.44

harmonic mean of precision and recall, where the F1 score is between 0 (worst) and 1 (best). This score is especially useful when dealing with unbalanced classifications as is the case in our dataset. We report two different kinds of F1 scores to adapt to the multiple label classification task: F1 weighted and F1 macro. F1 macro calculates the F1 score for each class separately and then takes the average of the scores. Thus, each class is given equal weight, regardless of its size. The weighted F1 scores weighs the category-specific F1 scores by the share of each category in the dataset. Accuracy refers to the share of correctly classified sentences and balanced accuracy is the mean of sensitivity and specificity averaged across categories. In addition, we report macro averages of Precision and Recall. We prefer macro averages as these are sensitive to changes in prediction quality in the less frequent categories.

Appendix D. Additional prompt engineering results

Figure A2 corresponds to Figure 4 inside the main text, pointing out the tradeoff between the number of sentences classified in one prompt regarding accuracy and the number of tokens used. Level 1 (topic) and level 2 (normative) appear to be more sensitive to the number of sentences included inside the prompt. We find highest accuracy for 10 sentences per prompt.

Figure A3 is the equivalent of Figure 3 in the main text, depicting how accuracy and stability of the classification vary with the temperature setting of ChatGPT for Level 1 (topic) and Level 2 (normative) classifications. As we found for level 3, accuracy is less stable with higher temperature and declines on average with higher temperature.

In Table A8, we present accuracy metrics of our final prompt with different messages. Leaving the default message "You are a helpful assistant" performs virtually identical to providing ChatGPT with a central bank context by assigning it the role of a research assistant at a central bank. Finally, a very elaborate system message giving ChatGPT the persona of a expert on central bank communication does, if anything, slightly lower the prediction accuracy.

TABLE A8. Validation metrics of system messages.

System Message	Accuracy	F1 (weighted)	F1 (macro)	Recall (macro)
You are a helpful assistant.	0.62	0.66	0.36	0.44
You are a research assistant at a central bank.	0.62	0.66	0.37	0.45
You are a distinguished expert on central bank communication. Through your thorough studies, having read countless speeches and other central bank documents, you are familiar with the language central bankers use and know how to interpret their statements. This expertise enables you to understand nuanced differences in central bank communications and accurately decode the sometimes hard to grasp messages conveyed inside their communication.	0.60	0.66	0.35	0.43

TABLE A9. Performance Metrics of different prompt configurations

	Minimal Instructions	**Final**	Detailed instructions
Accuracy	0.50	0.62	0.47
F1 (weighted)	0.57	0.66	0.55
F1 (macro)	0.30	0.36	0.29
Precision (macro)	0.29	0.34	0.31
Recall (macro)	0.44	0.44	0.40
Tokens used	106379	137193	16099
Prompt length	672	2433	3559

Note: The mid column indicates the validation metrics of our final prompt. The other prompts can be found in the Appendix.

FIGURE A2. Relationship between sentence count, accuracy and token usage

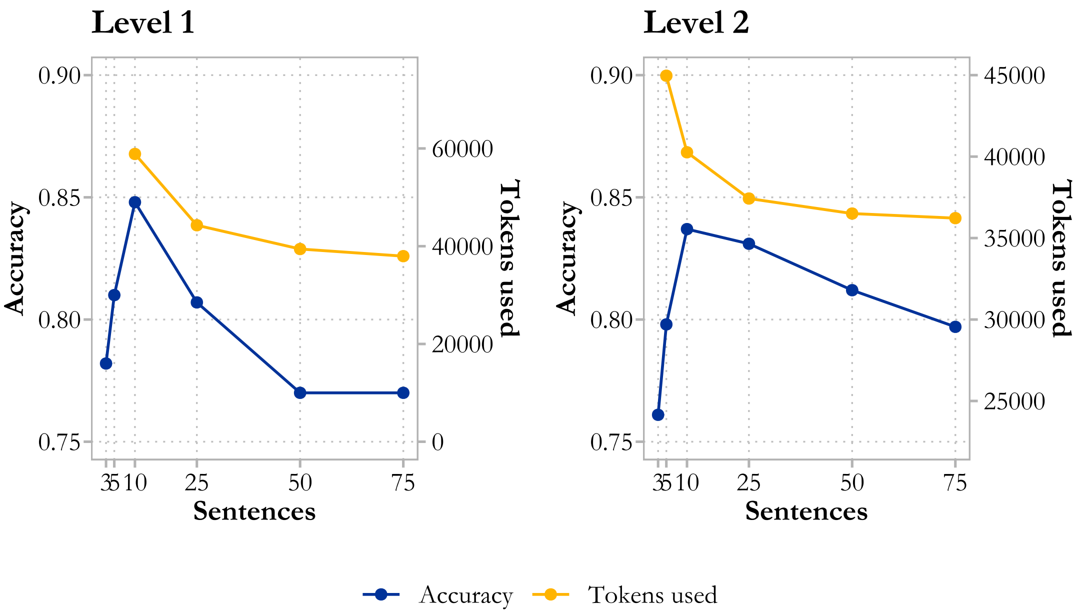
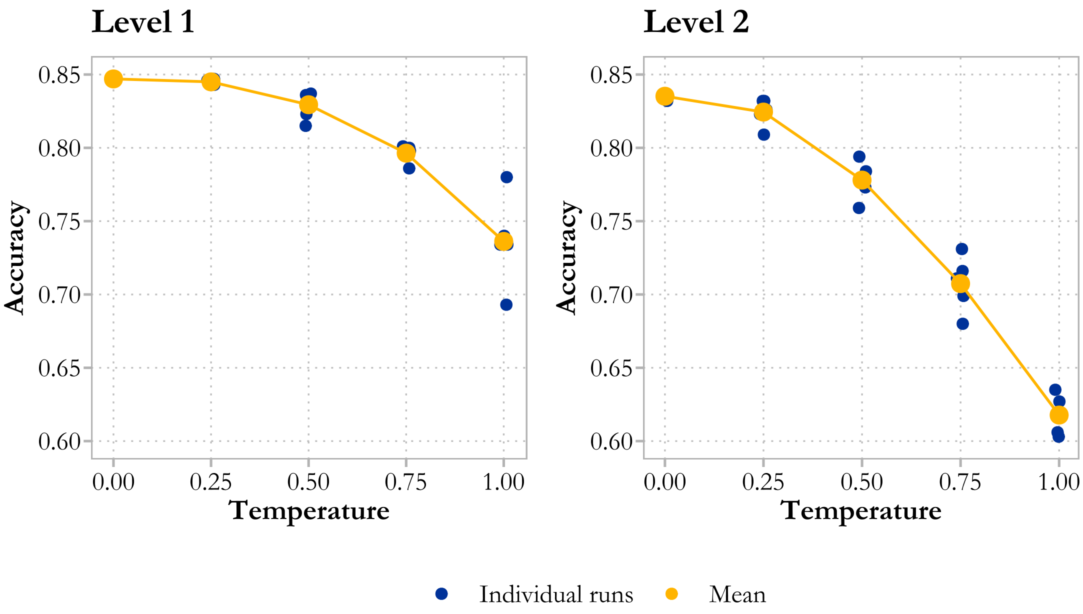


FIGURE A3. Variation in accuracy with different temperature settings



Appendix E. Additional variation in indices

Figures A5, A6, A7, A8 present further correlations of our dominance and coordination indicators vis-a-vis government bond spreads, polarization and purchasing power parity GDP per capita. Government bond spreads are measured as the difference of the 10 year government bond yield to the bond yield of the German government bond. The polarization indicator is constructed using the "v2cacamps_mean" variable from the VDEM dataset, whereby 0 and 1 map to low polarization and 2 and 3 map to high polarization. GDP figures are taken from the IMF's World Economic Outlook. Figure A4 shows the share of normative sentences over time for each of the country groups as defined in Table 6 of the main text.

FIGURE A4. Level 2 classification by country group

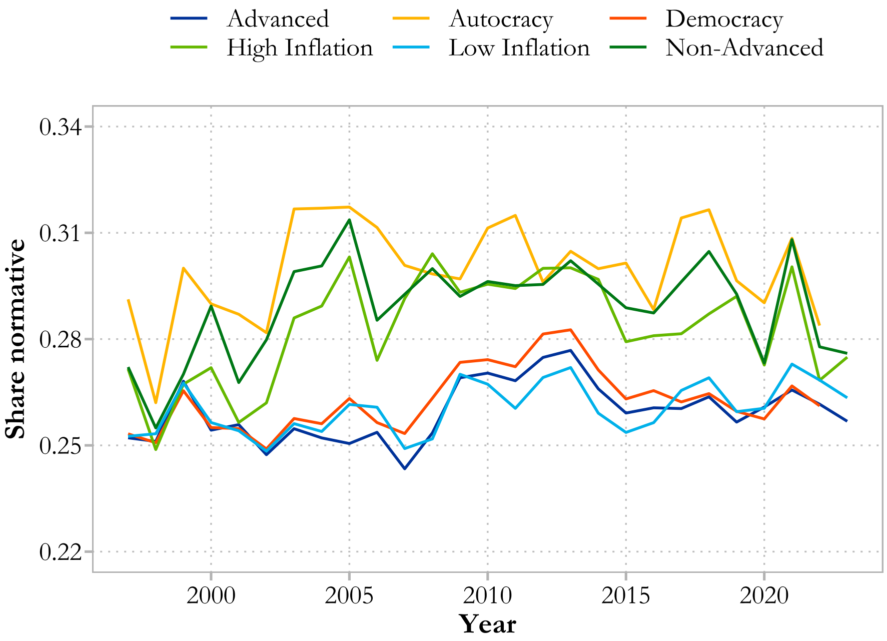
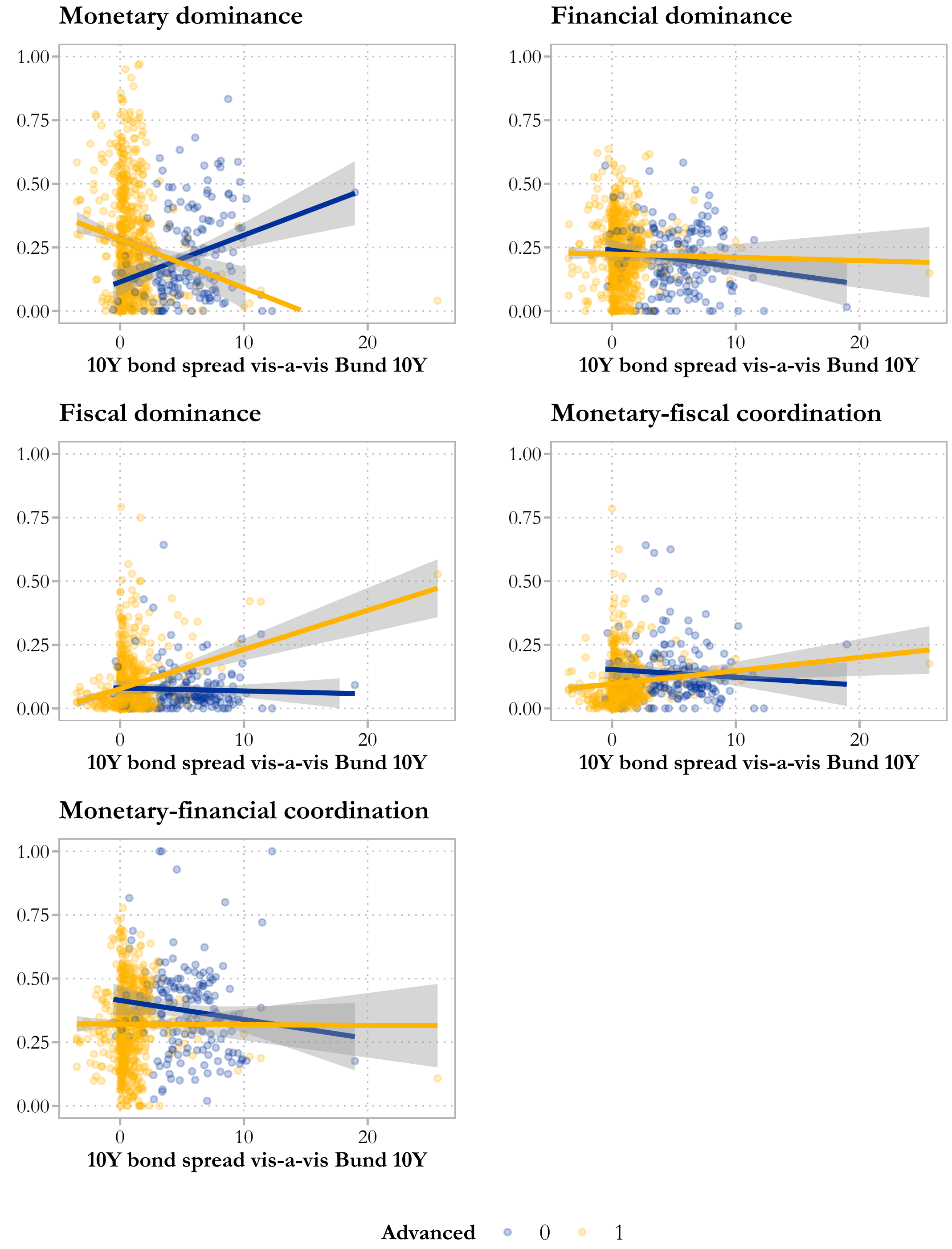
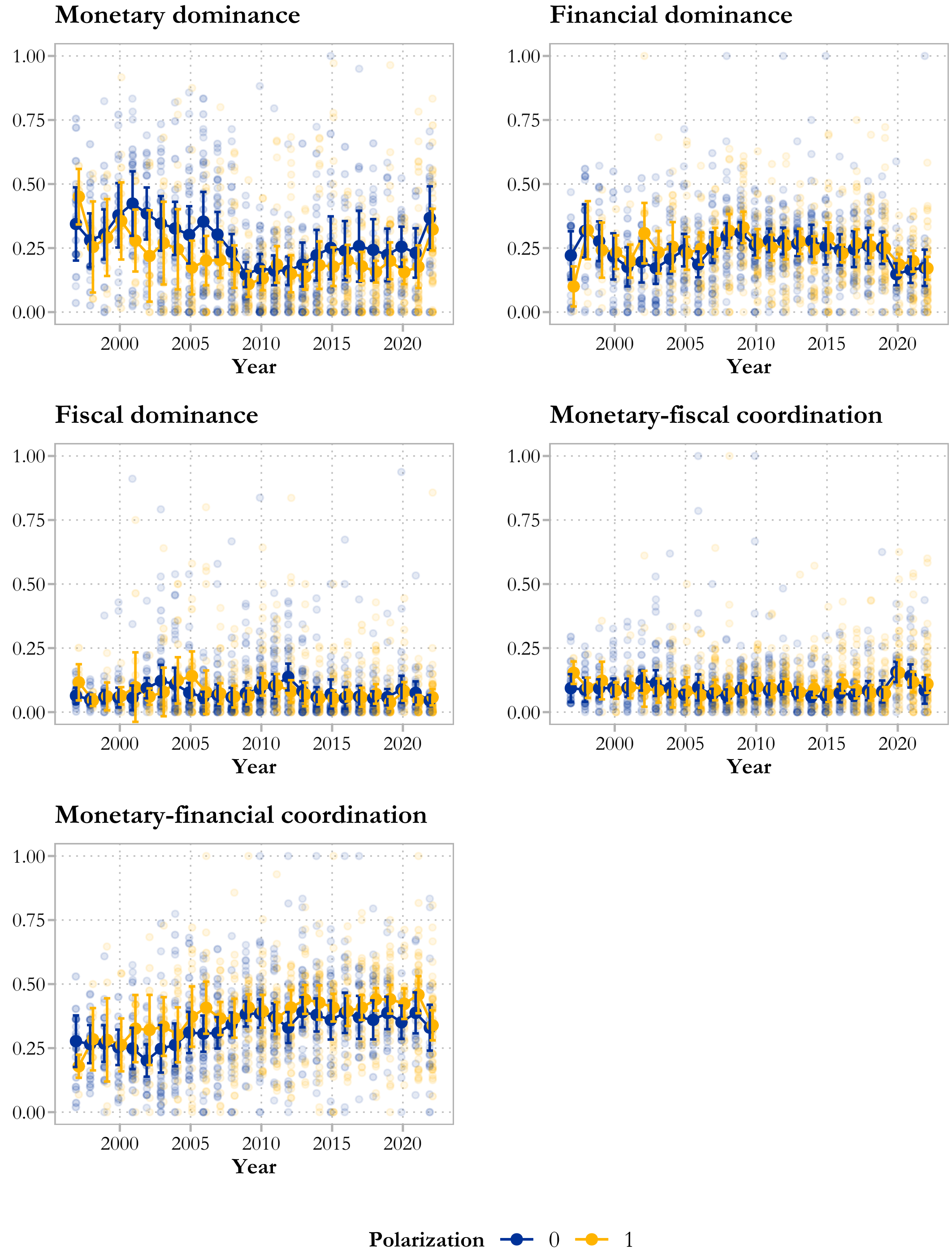


FIGURE A5. Scatter chart of correlation of bond spreads with coordination and dominances in advanced and non-advanced countries.



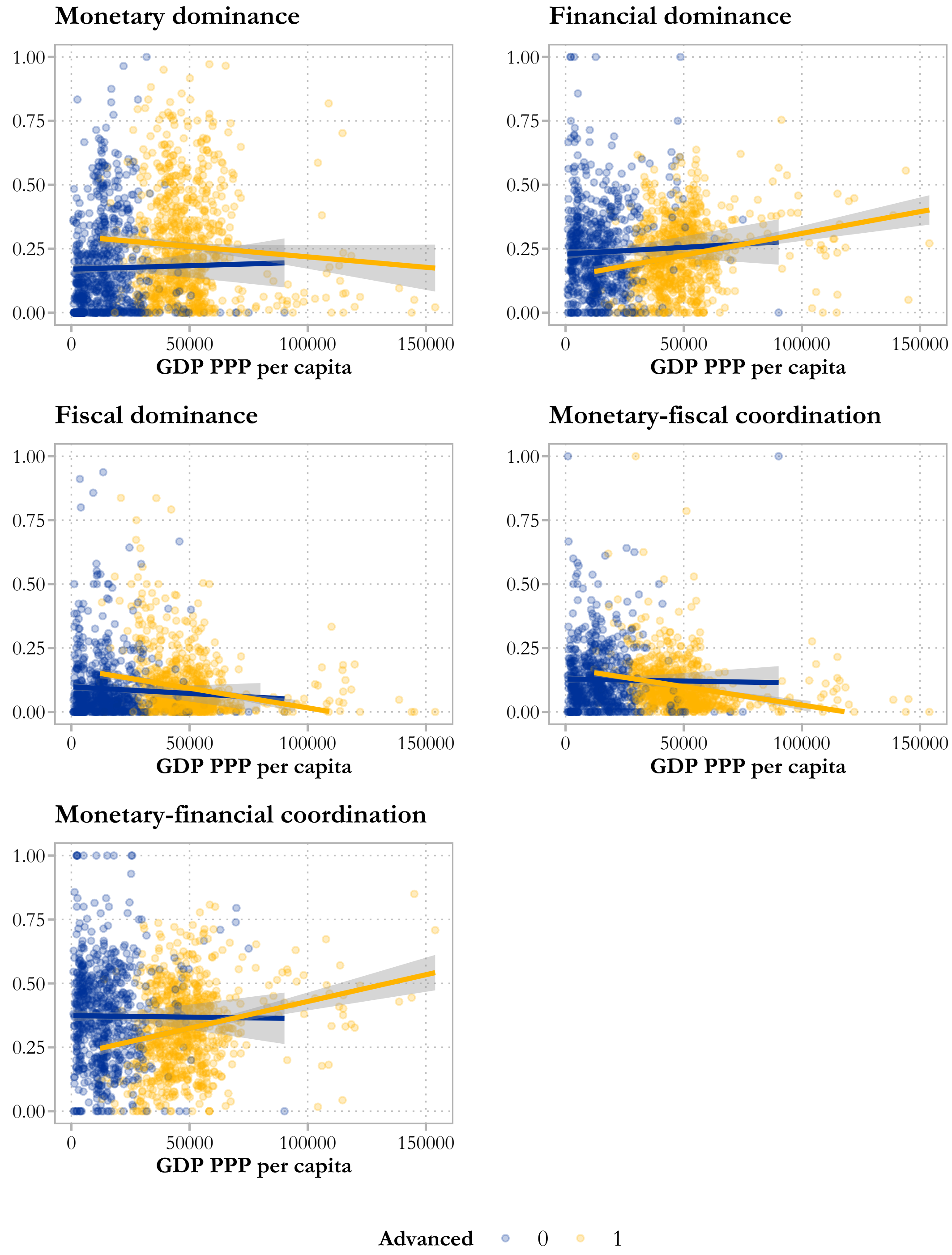
Note: Dots show central bank-year observations. The solid line indicates indicates a pooled linear fit by country group. The shaded region around the regression line is the 95% confidence band.

FIGURE A6. Jitter chart of the means of high and low polarization countries over time.



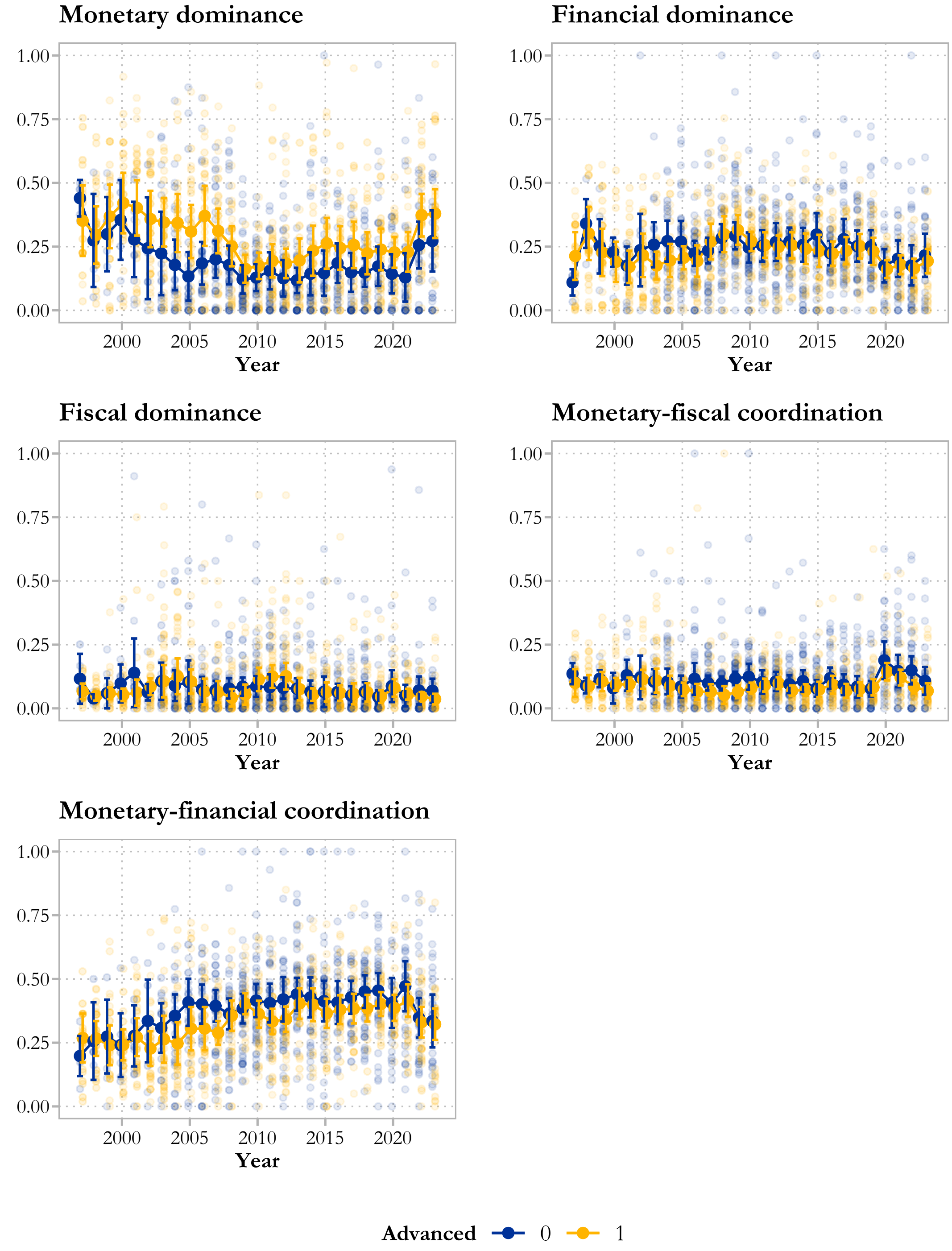
Note: Small dots show central bank-year observations. Observations with more speeches are less opaque. The solid line indicates the average of the category-average weighted by the number of speeches. Error bars indicate the 95% confidence interval of the weighted mean.

FIGURE A7. Scatter chart of correlation of GDP per capita (PPP) with coordination and dominances in advanced and non-advanced countries.



Note: Dots show central bank-year observations. The solid line indicates indicates a linear fit by country group. The shaded region around the regression line is the 95% confidence band

FIGURE A8. Jitter chart of the means of advanced and emerging and developing economies over time.



Note: Small dots show central-bank year observations. Observations with more speeches are less opaque. The solid line indicates the average of the category-average weighted by the number of speeches. Error bars indicate the 95% confidence interval of the weighted mean.

Appendix F. Prompt variations

F.1. Minimal Prompt

Classify excerpts from a central bank speech as one the following categories:

- "monetary dominance", "monetary-financial coordination", "monetary-fiscal coordination", "financial dominance" or "fiscal dominance" if the speaker suggests the presence thereof.
- Rely on your knowledge of what those categories mean. Classify in one of the coordination categories if there is no clear hierarchy. When there is no indication of a dominance or coordination relation with regard to fiscal authorities or financial markets classify as "none".

Classify each of the excerpts individually. Reply only with the number of the excerpt and the assigned label.
These are the excerpts:
1. <Excerpt 1>
2. <Excerpt 1>
…

F.2. Extended Prompt

Classify excerpts from a central banker speech as one of the following categories:

- Classify the excerpts as either financial, monetary, or fiscal dominance when there is a hierarchy between the central bank and another actor. Monetary dominance if the central banker is explicitly or implicitly prescribing a policy to others. Financial or fiscal dominance if the central banker explicitly or implicitly suggests that monetary policy will accommodate other policies.
- Classify it as fiscal-monetary coordination or monetary-financial coordination when there is not a clear hierarchy and there is a request to coordinate or cooperate.

More information:

- "none" if the central bank official:
    - when the excerpt is ambiguous to the extent that the context cannot be implied
    - this is also the default category
- "monetary dominance" if the central bank official:
    - emphasizes price stability above other objectives (e.g., inflation target is more important than financial stability or public debt sustainability)

  - suggests that governments should pursue prudent fiscal policy (e.g., fiscal consolidation, reduce public debt/deficit)
  - suggests to raise interest rates despite negative consequences on growth and employment
  - suggests a rules-based fiscal framework or new fiscal institutions
- “monetary-financial coordination” if the central bank official:
  - suggests to work together to create more efficient or better functioning financial markets (e.g., more market transparency)
  - suggests to facilitate liquidity of markets
  - suggests improving the deposit guarantee funds together with financial institutions
- “monetary-fiscal coordination” if the central bank official:
  - suggest to introduce a policy-mix of monetary and fiscal policy or tighter cooperation regarding fiscal and monetary policy
  - suggests to provide governments with additional liquidity through loan facilities, accept greater range of securities as collateral for the loans or loans with longer-than-usual maturities.
  - suggests in liaison with governments to use a combination of monetary easing, fiscal expansion, and targeted credit support.
  - suggests to ensure effective transmission of monetary policy to public and private spending.
- "financial dominance" if the central bank official:
  - suggests that the central bank should support the financial markets (e.g., provide liquidity support, recapitalize or bail out banks) regardless of inflation
  - suggests not to tighten the monetary policy stance (e.g. not to raise interest rates) if this threatens financial stability
  - suggests that financial market stress is making monetary policy more difficult
  - is concerned about negative feedback effects from the financial markets (e.g. doom loop, contagion)
  - suggests that the central bank will inject liquidity via non-bank financial intermediaries or repo markets, will bail out or recapitalise banks
- "fiscal dominance" if the central bank official:
  - is accommodating to government policies (e.g., suggests to bring down sovereign bond spreads to help reach sustainable growth).
  - suggests monetary financing (e.g., financing the government deficit)

- suggests to intervene in the sovereign bond market to stabilise it prioritizes government debt sustainability over price stability (e.g. a situation where fiscal policy is not sustainable forcing the central bank to prioritise the government's solvency above its own objectives)

Classify each of the excerpts individually. Reply only with the number of the excerpt and the assigned label.

These are the excerpts:

1. <Excerpt 1>
2. <Excerpt 1>

...